Attorney Docket No. 0826.1726

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Yoshio NAKAO

Application No.: 09/862,437 Group Art Unit: 2654

Filed: May 23, 2001 Examiner: Abdelalil Serrou

For: APPARATUS FOR RADING A PLURALITY OF DOCUMENTS AND A METHOD THEREOF

SUBMISSION OF VERIFIED ENGLISH TRANSLATION OF FOREIGN PRIORITY APPLICATION

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

Attached is the English translation of Japanese Patent Application No. 2000-290886, filed September 25, 2000. It is respectfully requested that the attached English translation be made of record in the above-identified application.

If any additional fees are required in connection with the filing of this document, please charge Deposit Account No. 19-3935.

Respectfully submitted,

STAAS & HALSEY LLP

Date: 10/11/05

By: Richard A. Gollhofer
Richard A. Gollhofer
Registration No. 31,106

1201 New York Ave, N.W., Suite 700
Washington, D.C. 20005
Telephone: (202) 434-1500
Facsimile: (202) 434-1501

VERIFICATION

I, Tohru IGARASHI, residing at Kanagawa, Japan, state: that I know well both the Japanese and English languages; that I translated, from Japanese into English, the priority document as filed in the U.S. Patent Application No. 09/862,437, filed on May 23, 2001; and that the attached English translation is a true and accurate translation to the best of my knowledge and belief.

Dated: October 4, 2005

Tohru Igarashi

Tohru IGARASHI

[Document Name]	Patent Application
[Reference No.]	0051731
[Filing Date]	September 25, 2000
[Addressee]	Commissioner, Patent Office
[Int'l Patent Classification]	G06F 17/20
[Title of Invention]	Apparatus for Reading a Plurality of Documents and a Method thereof
[Number of Claims]	9
[Inventor]	
[Address or Residence]	c/o FUJITSU LIMITED 1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki-shi, Kanagawa
[Name]	Yoshio NAKAO
[Patent Applicant]	
[Identifying No.]	000005223
[Name]	FUJITSU LIMITED
[Agent]	
[Identifying Number]	100074099
[Address or Residence]	3rd Fl., Nibancho Bldg., 8-20, Nibancho, Chiyoda-ku, Tokyo
[Attorney]	
[Name or Title]	Yoshiyuki OSUGA
[Telephone No.]	03-3238-0031
[Agent Appointed]	
[Identifying No.]	100067987
[Address or Residence]	503, 25-28, Kitaterao 7-chome, Tsurumi-ku, Yokohama-shi, Kanagawa
[Attorney]	
[Name or Title]	Akira KUKIMOTO
[Telephone Number]	045-573-3683
[Fee Designation]	
[Pre-payment Reg. No.]	012542
[Payment Amount]	JPY21,000

[Index of Submitted Article]		
[Article Name]	Specification	1
[Article Name]	Drawings	1
[Article Name]	Abstract	1
[General Power of Attorney No.]	9705047	
[Necessity of Proof]	Yes	

[Document Name] Specification

[Title of the Invention] Apparatus for Reading a Plurality of Documents and a Method thereof

[What is claimed is:]

1. A document reading apparatus presenting a plurality of documents designated as reading documents by a user, comprising:

a thematic hierarchy recognition device recognizing a thematic hierarchy of each of the plurality of documents;

a topic extracting device extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

a topic relation presentation device taking out a description part corresponding to the extracted topic from each of the plurality of documents and outputting the taken-out description parts.

2. The document reading apparatus according to claim 1, wherein regarding a topic set that consists of topics of various grading in the recognized thematic hierarchies, the topic extracting device calculates a relevance score between topics of the topic set based on lexical similarity of description parts corresponding to each topic of the topic set, and extracts a topic set having a relevance score equal to

or more than a threshold that is set based on inclusive relationship of topics.

3. The document reading apparatus according to claim 1, wherein the topic relation presentation device presents the taken-out description parts side by side.

4. The document reading apparatus according to claim 3, wherein the topic relation presentation device presents the related parts and original documents in two windows, one of the windows including the related parts side by side and the other including the original documents side by side.

5. The document reading apparatus according to claim 3, wherein the topic relation presentation device presents summaries of the related parts.

6. The document reading apparatus according to claim 3, wherein the topic relation presentation device presents a plurality of thematic hierarchies corresponding to the plurality of documents and a correspondence relationship between the plurality of thematic hierarchies based on the plurality of common topics in a drawing, and presents a designated part of the plurality of documents in accordance with an instruction from the user given on the drawing.

7. The document reading apparatus according to claim 1, wherein the topic relation presentation device sets

one document among the plurality of documents as a reference document, generates a new integrated document by merging the contents of the reference document with description parts of another document related to the reference document, and outputs the integrated document.

8. A computer-readable storage medium storing a program for a computer that presents a plurality of documents designated as reading documents by a user, the program causes the computer to perform:

recognizing a thematic hierarchy of each of the plurality of documents;

extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

extracting a description part corresponding to the extracted topic from each of the plurality documents and outputting the extracted description part.

9. A document presenting method of presenting a plurality of documents designated as reading documents by a user, comprising:

recognizing a thematic hierarchy of each of the plurality of documents;

extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

extracting a description part corresponding to the extracted topic from each of the plurality documents and outputting the taken-out description parts.

[Detailed Explanation of the Invention]

[0001]

[Field of the Invention]

The present invention relates to an apparatus for reading a machine-readable document on the screen of a computer, and a method thereof. Especially, the present invention intends to support the comparative reading work of the related documents by presenting the related passages across the documents to be compared in a form of easily understanding.

[0002]

[Prior Art Technology]

If there is a technology for presenting the related part in the plurality of documents easy to read when a user reads a plurality of related documents while comparing them to each other, the efficiency of the comparison work can be improved. For example, when a user reads a plurality of survey reports from researchers in each area in order to make a summary report on the actual situations about a specific survey item of a plurality of areas or when a user reads and compares a question document and its reply document, a technology for support the comparison of related parts is demanded. As for representative articles regarding the multi-document comparison support, following seven pieces are cited:

[1] Christine M. Neuwirth and David S. Kaufer, "The Role of External Representations in the Writing Process: Implications for the Design of Hypertext-based Writing Tools", In Proc. of Hypertext '89, pp. 319-341, the Association for Computing Machinery (Nov. 1989).

[2] Nobuyuki Omori, Jun Okamura, Tatsunori Mori, and Hiroshi Nakagawa, "Hypertextixation of a Relation Manual Group Using tf. idf Method", Information processing academy research report FI-47-8/NL-121-16, Information Processing Academy (Sep. 19979.

[3] Gerard Salton, Amit Singhal, Chris Buckley, and Mandar Mitra, "Automatic Text Decomposition Uusing Text Segments and Text Themes", In Proc. of Hypertext '96, pp. 53-65, the Association for Computing Machinery (Mar. 1996).

[4] Inderjeet Mani and Eric Bloedorn, "Summarizing Similarities and Differences among Related document", Chapter 23, pp. 357-379. The MIT Press, London (1999). (reprint of Information Processing and Management, Vol. 1, No.1. pp. 1-23(1999)).

[5] Japanese Patent Laid-open Publication No. 7-325,827

[6] Japanese Patent Laid-open Publication 2000-57,152 (P2000-57152A)

[7] Japanese Patent Laid-open Publication No. 11-39,334

Among these, the document [1] proposes an interface (screen) called "Synthesis Grid" which summarizes the similarities and differences across related articles in an author-proposition table.

[0003]

Also, as for the conventional technology for extracting the related parts across documents, the technology that sets a hyperlink across the related parts of different documents with a clue of the appearance of the same vocabulary has been known. For example, the article [2] shows the technology for setting a hyperlink between a pair of document segments that show high lexical similarity. The articles [5] and [6] show the technology for setting a hyperlink across the related parts among documents where the same keyword appears.

[0004]

In addition, the article [3] shows the technology for extracting the related parts in a single document by detecting the paragraph group having a high lexical similarity. Also, the article [4] shows a method for discovering topic-related textual regions based on coreference relations using spreading activation through coreference of adjacency word links.

[0005]

As for the technology for presenting similarities and differences of a plurality of related documents, the article [7] shows a multi-document presentation method that distinguishes the information commonly included in a plurality of documents from the other information. The method displays the whole contents of one selected article with highlighting (hatching) common information, and supplements unique information about remaining articles.

[0006]

[Problems to be Solved by the Invention]

However, there are the following two problems in the above-mentioned conventional technology.

The first problem is that it is difficult to detect appropriate related parts regarding a topic that differs in grading since a unit for recognizing the related parts is fixed. Namely, since in the above-mentioned prior art, the unit of the comparison is fixed to one of a section, a paragraph and a sentence (or the appearance position of a word), only parts in the comparison unit, such as a section, a paragraph and the like, can be basically detected.

[0007]

Therefore, for example, if a part consisting of two paragraphs in the first reading document is related to a part consisting of several or more paragraphs in the second reading document, it is difficult to appropriately extract related parts in the comparable form. In order to realize this, another measure must be taken, for example, parts detected as related parts must be merged.

[0008]

The second problem is that the relationship between an aggregate of related parts regarding a certain topic and either another aggregate of those regarding a different topic or the whole original document cannot be clearly expressed. For example, when long documents with complex topics are compared, related pars regarding a plurality of topics are sometimes intertwined and detected.

[0009]

In such a case, not only an aggregate of related parts across documents regarding each topic must be mutually compared, but also must be examined in detail taking into consideration the mutual relationship between a plurality of topics common among the documents, the context in which each related part appears and the like. In this case, it is preferable to be able to take a look at a plurality of aggregates of related parts and to easily refer to the periphery of each related part. However, the above-mentioned prior art cannot realize such a function.

[0010]

The first object of the present invention is to provide a document reading apparatus for extracting and presenting an appropriate related part regarding a topic that differs in grading for each document, and a method thereof. The second object of the present invention is to provide a document reading apparatus for presenting many aggregates of related parts across documents regarding a plurality of topics in a form easily compared and analyzed, and a method thereof.

[0011]

[Means for Solving the Problems]

Fig. 1 is a block diagram showing the principle of the document reading apparatus of the present invention. The document reading apparatus shown in Fig. 1 comprises a thematic hierarchy recognition device 1, a topic extracting device 2 and a topic relation presentation device 3. The apparatus presents a plurality of documents that are designated as a reading object to a user, and supports the comparison work of those documents.

[0012]

The thematic hierarchy recognition device 1 recognizes the respective thematic hierarchies of a plurality of aggregates of documents to be read. Here, the thematic hierarchy means that a plurality of topics constituting a document constitutes a hierarchy composed of two or more paragraphs. This hierarchy corresponds to the inclusive relationship of topics such that each of the plurality of aggregates of large topics constituting a document includes a plurality of aggregates of small topics and each of the plurality of aggregates of small topics includes a plurality of aggregates of smaller topics.

[0013]

The topic extracting device 2 extracts topics that commonly appears in a plurality of the documents to be read, based on the recognized thematic hierarchies. In this case, a plurality of thematic hierarchies that individually correspond to a plurality of documents are compared, and the combination of topics having strong relevance is extracted to be output as a common topic across a plurality of documents. For example, if the first and second thematic hierarchies are obtained from documents D1 and D2, the relevance score of each pair of nodes (topics) from the first and second thematic hierarchies is calculated, and topic pairs with a high relevance score are extracted to be outputted as common topics.

[0014]

The topic relation presentation device 3 extracts a pair of description parts corresponding to the extracted common topics from each document to be read. It this case, the extracted description parts are outputted as related parts across the plurality of documents to be read.

[0015]

In this way, the document reading apparatus detects topics of various grading (sizes) that are included in each document to be read using the thematic hierarchy recognition device 1. The apparatus then extracts common topics across the documents from the detected topics using the topic extracting device 2. Finally, the apparatus extracts and outputs an aggregate of description parts of the documents for each topic that the topic extracting device 2 extracts using the topic relation presentation device 3.

[0016]

By inclusively checking the relevance of the topics while using each of the topics of various grading that are included in the documents to be read as a unit, the correspondence relationship of the description parts with different sizes can be detected. For example, even when a part consisting of two paragraphs of the document D1 corresponds to a part consisting of several or more paragraphs of the document D2 as an aggregate, an appropriate related pars can be extracted.

[0017]

Furthermore, the document reading apparatus of Fig. 1 has the following various functions:

The topic extracting device 2 obtains the relevance score between topics by the lexical similarity of a passage in the document corresponding to each topic, and selects a pair of topics as a common topic (group) by the threshold that is set based on the inclusive relationship of topics. For example, a pair of topics A and B in an upper layer with a relevance score R1 is output as a common topic, only when none of the smaller topics included in topic A or topic B shows a relevance score equal to or more than R1.

[0018]

In this way, the output of an inappropriate related part is restrained, so that the related pars can be more efficiently output. For example, if there is an aggregate of topics consisting of a plurality of paragraphs in each of two documents to be read and they are related, a part of the paragraphs constituting those aggregate is also sometimes related in parallel as a topic of some grading.

[0019]

Specifically, if relationship can be detected between the aggregate of two paragraphs of the first and second paragraphs of document D1 and the aggregate of two paragraphs of the first and second paragraphs, sometimes relationship can be detected between the first paragraphs of documents D1 and D2 or between the second paragraphs of documents D1 and D2. In such a case, the document reading apparatus can appropriately determines whether the related part is output as relationship between the aggregates or as relationship between the respective paragraphs, so that redundant output can be suppressed.

[0020]

Furthermore, the topic relation presentation device 3 groups related parts by each common topic and presents the grouped part side by side. In this way, a user can read the corresponding parts regarding each topic while contrasting them, even when a plurality of common topics are detected.

[0021]

Furthermore, the topic relation presentation device 3 can summarize and output the contents of each related part. In this way, a user can take a look at the whole list of related parts, even when many common topics are detected.

[0022]

Furthermore, the topic relation presentation device 3 can link and present a part of the original document corresponding to each related part. For example, the button (hyper link or the like) for original document reference is provided in each related part in a window, and the corresponding part of the original document is presented in another window in accordance with the request made by the button. In this way, a user can examine the contents of each related part in the context where it appears.

[0023]

The topic relation presentation device 3 also presents a drawing showing the thematic hierarchy of the document to be read, and presents the corresponding parts of the document in accordance with the designation of a user on the screen. For example, the device presents two tree-structure graphs of a thematic hierarchy in which a node depicts a topic or presents a common topic as an arc between the nodes, thereby receiving a user's request. When a user designates an arc, the device presents a related part corresponding to the arc in another window. When a node is designated, the device similarly presents a part corresponding to the node.

[0024]

In this way, a user can examine a related part while referring to the context or the like as required using the topic configuration of the entire document as a clue, so that a plurality of documents can be more efficiently compared and read.

[0025]

Further, the topic relation presentation device 3 generates and presents a new integrated document by using one document as a reference document and extracting related parts from another documents regarding a common topic. In this way, a user can effectively generates an integrated document, such as a report obtained by merging a plurality of documents, or the like.

[0026]

[Preferred Embodiments]

The preferred embodiments of the present invention are described in detail below with reference to the drawings.

The present invention relates to a function to comprehensively present the similarities and difference among documents, and realizes the function to the most advance automation using the present technology. More particularly, the present invention realizes a function to contrast and present related parts in a plurality of related documents, utilizing an automatic extraction technology of related parts in the documents.

[0027]

Fig. 2 shows the basic configuration of the document reading apparatus of the present invention. The document reading apparatus 12 of Fig. 2 comprises an input unit 21, a tokenizer 22, a machine readable dictionary 24, a thematic hierarchy detector 25, a topic extractor 27, and an output unit 28.

[0028]

The thematic hierarchy recognition device 1, topic extracting device 2,and topic relation presentation device 3 of Fig. 1 correspond to the thematic hierarchy detector 25, the topic extractor 27, and the output unit 28 of Fig. 2, respectively.

[0029]

In Fig. 2, when a plurality of documents to be read

11 are input, the document reading apparatus 12 extracts related parts corresponding to a common topic across those documents, and presents the extracted related parts to a user 13.

[0030]

The input unit 21 reads a plurality of documents to be read 11, and sequentially transfers each document to the tokenizer 22. The tokenizer 22 linguistically analyzes each document 11 using a morphological analyzer 23,which is a sub-module, and extracts content words (e.g., noun, verbs, adjectives, adjectival verbs) in the document 11 and marks the corresponding parts of the document 11. At this time, the morphological analyzer 23 converts a sentence in the document 11 to a word list with parts of speech information in reference to the machine readable dictionary 24. The machine-readable dictionary 24 is a word dictionary for a morphological analysis, and describes the correspondence relationship between the notation character string and the information about the parts of speech and inflection (conjugation) type of a word.

[0031]

The thematic hierarchy detector 25 receives a plurality of documents to be read 11 with the marks of content words, recognizes the thematic hierarchy of each document 11, and outputs it. First of all, the thematic hierarchy detector 25 automatically recognizes an aggregate of topics of various grading (sizes) in the document, using a thematic boundary detector 26. Here, an aggregate of topics means a part of the document that describes a common topic. The thematic hierarchy detector 25 relates an aggregate of large topics to an aggregate of small topics, generates thematic hierarchy data, and outputs it.

[0032]

The thematic boundary detector 26 recognizes a section with a low lexical cohesion score as a candidate section of a thematic boundary. The lexical cohesion score indicates the strength of cohesion concerning a vocabulary in the vicinity of each position in the document. For example, it can be obtained from the similarity of the vocabulary that appears in a window of a certain width that is set before and after each position.

[0033]

The topic extractor 27 receives a plurality of thematic hierarchies each corresponding to each of a plurality of documents to be read 11 from the thematic hierarchy detector 25, detects a topic that commonly appears in two or more documents, and outputs a list of the common topics.

[0034]

The output unit 28 extracts the description parts corresponding to each of the common topics that are extracted by the topic extractor 27, correlates these related parts, and presents the correlated part to a user 13.

[0035]

The document reading apparatus 12 of Fig. 2 can be configured, for example, using the information processor (computer) as shown in Fig. 3. The information processor of Fig. 3 comprises an output device 41, an input device 42, a CPU (central processing unit) 43, a network connection device 44, a medium driving device 45, an auxiliary storage device 46 and memory (main memory) 47, which are mutually connected by a bus 48.

[0036]

The memory 47 includes, for example, a ROM (read only memory), a RAM (random access memory), etc., and stores the program and data that are used for a document reading process. Here, the input unit 21, tokenizer 22, morphological analyzer 23, thematic hierarchy detector 25, thematic boundary detector 26, topic extractor 27, and output unit 28 are stored as a program module. The CPU 43 performs a required process by running the program utilizing the memory 47.

[0037]

The output device 41 is, for example, a display, a printer, or the like. It is used for an inquiry to a user 13 and the output of the document to be read 11, the processing result, etc. The input device 42 is, for example, a keyboard, a pointing device, a touch panel, a scanner or the like. It is used to input instructions from the user 13, and the document to be read 11.

[0038]

The auxiliary storage device 46 is, for example, a magnetic disk device, an optical disk device, a magneto-optical device, or the like, and stores the information of the document to be read 11, machine readable dictionary 24, etc. The information processor stores the above-mentioned program and data in the auxiliary storage 46, and it loads them into the memory 47 to be used, as occasion demands.

[0039]

The medium driving device 45 drives a portable storage medium 49, and accesses its record contents. As for the portable storage medium 49, an optional computer-readable storage medium such as a memory card, a floppy disk, a CD-ROM (compact disk read only memory), an optical disk, a magneto-optical disk, or the like is used. The user 13 stores the above-mentioned program and data in the portable storage medium 49, loads them into the memory 47 to be used, as occasion demands.

[0040]

The network connection device 44 communicates with an external device through an arbitrary network such as a LAN (local area network), etc., and performs the data conversion associated with the communication. The information processor also receives the above-mentioned program and data from another device such as a server, etc., through the network connection device 44, and loads them into the memory 47 to be used as occasion demands.

[0041]

Fig. 4 shows a computer-readable storage medium that can supply a program and data to the information processor of Fig. 3. The program and data that are stored in the database 51 of a portable storage medium 49 and a server 50 are loaded into the memory 47. Then, the CPU 43 runs the program using the data, and performs a required process. At this time, the server 50 generates a carrier signal for carrying the program and data, and transmits the signal to the information processor through an arbitrary transmission medium on the network.

[0042]

Next, the operation of each module of the document reading apparatus 12 that is shown in Fig. 2 is described in more detail using a specific example.

As for an example of the documents to be read, the representative question made by Hiroko Mizushima, Diet member (the first document to be read) and the answer of the Prime Minister to the question (the second document to be read) that are respectively extracted as one document from "The Minutes No.2 of The 149th Plenary Session of the House of Representatives" (on July 31, 2000) are used. The representative question of the House of Representatives is advanced in such a way that the prime minister and related minister answers the questions, after the Diet member who represents a political party asks questions about several items in a bundle. In this representative question, the total eight items are questioned regarding six problems of child education, civil law revision, Diet operation, harmful information, infant medical treatment, and annual expenditure payment method.

[0043]

Fig. 5 shows the leading part of the first document to be read that is extracted from the representative question part. In Fig. 5, since the underlined part, specifically, the name of the Diet member who asks the question and the supplementary information regarding the proceeding progress that is parenthesized) are not the contents of a representative question, the following processes are performed after they are removed. As for the second document to be read that is obtained by extracting the prime minister's answer part, the name of the prime minister, and the supplementary information that is parenthesized are similarly removed.

[0044]

Fig. 6 is a flowchart showing the word recognition process by the tokenizer 22. First of all, the tokenizer 22 applies a morphological analysis to each document to be read, and generates a word list with the parts of speech (step S11). Then, the tokenizer 22 recognizes content words (nouns, verbs, adjectives, adjectival verbs, etc.) using part of speech information stored in the word list as a clue and marks the positions of the document corresponding the content words (step S12), and terminates the process. Fig. 7 shows the processing result of the tokenizer 22 for the document part of Fig. 5.

[0045]

In step S11 of Fig. 6, the morphological analyzer 23 performs a morphological analysis process as shown in Fig. 8. First of all, the morphological analyzer 23 clears the word list (step S21), attempts to extract a sentence from the beginning of the (remaining) document using a period or the like as a clue (step S22), and determines whether the sentence can be extracted (step S23).

[0046]

If the sentence can be extracted, then the morphological analyzer 23 obtains word candidates that are possibly included in the sentence in reference to the machine readable dictionary 24 (step S24). In the case of Japanese, since a word boundary is not formally clarified as shown in Fig. 7, all the words corresponding to the character sub-strings that are included in the sentence are obtained as candidates. If a sentence, for example, "東京都は大都市だ" is extracted, all the character sub-strings that are included in this sentence become word candidates as shown in Fig. 9.

[0047]

In the case of English, on the contrary, since words are explicitly separated by spaces, it becomes the main function of the morphological analysis to determine the parts of speech for each word corresponding to a character string separated by spaces. For example, if a sentence "Tokyo is the Japanese capital." is extracted, the root forms and parts of speech of five words that are explicitly included in this sentence are determined, as shown in Fig. 10.

[0048]

Then, the morphological analyzer 23 selects an adequate series of words from a viewpoint of the adjacency probability at the level of the parts of speech (step S25), adds information about the part of speech and the appearance position of each word to the selected series of words and adds the selected series of words to the word list in order of appearance (step S26). Then, the morphological analyzer 23 attempts to extract a subsequent sentence (step S27), and repeats the processes in and after step S23. When no sentence can be extracted in step S23, the process terminates.

[0049]

In the word recognition result of Fig. 10, the part that is parenthesized in ink is the content word that the morphological analyzer 23 recognizes. If the content word is a conjugative word (verb and adjective), the part before a slash (/) in the parentheses in ink indicates the root of word and a part after the slash (/) indicates the conjugative ending of the predicative form. Although this information is used to distinguish the word in a subsequent process, the subsequent process can also be performed using information such as the parts of speech and conjugation type, instead of this information. In short, arbitrary identification information can be used as long as the information distinguishes the word that cannot be distinguished by only the root of the word, for example, "い/う " and " い/る".

[0050]

Further, in step S25, various methods of evaluating the validity of the arrangement of words have been known as morphologic analysis methods, and an arbitrary method can be used. For example, the method of evaluating the validity of the arrangement of words using the appearance probability that is estimated by training data is reported in the following reference literatures [8], [9] and [10].

[8] Eugene Charniak, "Hidden Markov Models and Two Applications", In Statistical Language Learning, chapter 3, pp.37-73, The MIT Press (1993).

[9] Masaaki Nagata, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-best Search Algorithm", In Proc. of Information Processing Society of Japan, NL-101-10, IPS of Japan (May 1944)

[10] Masaaki Nagata, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A*

N-best Search Algorithm", In Proc. of COLING "94, pp. 201-207 (Aug. 1994).

In the example of Fig. 7, although the tokenizer 22 extracts all the content words, the extraction target can also be limited to nouns. Furthermore, when processing an English document, words can also be extracted after of all the words separated by open spaces, the vocabulary that appears everywhere irrespective of a topic (function word such as articles, prepositions, etc. or the words which appear too frequently) are removed, instead of performing the morphological analysis process. Such a process can be easily realized by preparing a stop word list that stores the function word or the word that appears with the high frequency instead of the machine readable dictionary 24.

[0051]

Next, the process of the thematic hierarchy detector 25 is described. In the present embodiment, an aggregate of topics is recognized based on the technology shown in the Japanese Patent Laid-open Publication No. 11-272,699 "Document Summarizing Apparatus and its Method" that is the prior application of the present invention. In this method, the hierarchical structure of a topic is recognized in the following procedures.

1. Estimation of the section of a thematic boundary position

A position where a thematic boundary might exist is obtained as the thematic boundary candidate section, on the basis of the cohesion score that is calculated using a certain window width. Then, this process is repeated for a plurality of window widths that differ in size, and the thematic boundary candidate section is obtained for each size of a topic, ranging from boundaries showing the gap of large topics to boundaries showing the gap of small topics.

2. Recognition of the hierarchical relation of topics

The thematic boundary candidate sections that are obtained using the different window widths are integrated, and the hierarchical structure of topics and thematic boundaries are determined.

[0052]

Fig. 11 is a flowchart showing the thematic hierarchy recognition process performed by the thematic hierarchy detector 25. The thematic hierarchy detector 25 first receives three parameters of the biggest window width w1, the minimum window width w_min, a window width ratio r from a user (step S41), and obtains a set W of window widths to measure the cohesion score (step S42). The set W of window widths is generated by collecting terms equal to or more than w_min from the geometric series the first term of which is w1 and the common ratio of which is 1/r.

[0053]

At this time, it is sufficient in practical use to give about 1/2 to 1/4 of the size of the entire document for the biggest window width w1 in W, to give the words equivalent to a paragraph (for example 40 words) for the minimum window width w_min, and to give 2 for the window width ratio r. In the following description, the value of w1=320 (word), w_min=40 (word), and r=2 are used.

[0054]

Then, the thematic hierarchy detector 25 calculates the cohesion score of each position in the document for each window width in W on the basis of the document the content word of which is marked as shown in Fig. 7, and records it as a cohesion score series (step S43).

[0055]

Here, the vocabularies that appear in two windows set before and after each position (reference point) of the document are first compared. A value that becomes bigger as the number of common vocabularies increases is calculated, and the calculated value is set as the cohesion score at the position. Then, the calculation of the cohesion score is repeated while shifting the position of the window at a fixed-width (tic) intervals from the leading part of the document toward the end. The calculated cohesion score is recorded as a series from the leading part of the document to the end.

[0056]

Any interval width tic is acceptable if the value is smaller than window width. Here, 1/8 of the window width is used in consideration of the processing efficiency. The value tic can also be designated by a user.

[0057]

Various methods are available as the calculation method of a cohesion score. In the following description, cosine measure, which has been widely used as the scale of similarity in the field of information retrieval or the like is used. The cosine measure is obtained by the following equation.

[0058]

[Mathematics 1]

$$\mathrm{sim}(b_l, b_r) = \frac{\sum_t w_{t,\,b_l} w_{t,\,b_r}}{\sqrt{\sum_t w^2_{t,\,b_l} \sum_t w^2_{t,\,b_r}}} \qquad (1)$$

[0059]

Here, b_l and b_r represent the part of the document that is included in the left window (window on the side of the leading part of the document) and the right window (window on the side of the end part of the document), respectively. $W_{t,bl}$ and $w_{t,br}$ represent the appearance frequency of word t that appears in the left window and right window, respectively. Also, Σ_t of the right side of the equation (1) represents the sum of the number of words t.

[0060]

The similarity of the equation (1) increases (the maximum 1) as the number of common vocabularies that are included in the right and left windows increases, while the similarity becomes 0 when there is no common vocabulary. Namely, a part with a high similarity score is expected to describe a common or similar topic. Conversely, a part with a low score is expected to contain a thematic boundary.

[0061]

Next, Fig. 12 shows an example of the series of the cohesion scores that are recorded in step S43. In Fig. 12, 1/4 of the window width w is used as interval width tic, to make the explanation simple. Document areas a1 through a11 are areas having the fixed width corresponding to the interval width tic. Also, c1 shows the cohesion score of window width w, which is calculated by using the boundary between a4 and a5 in the document as a reference point. In other words, c1 is the cohesion score calculated by using document areas a1 through a4 as the range of the left window and using document areas a5 through a8 as the range of the right window.

[0062]

The next c2 represents the cohesion score of the window width w that is calculated by shifting the window to the right by tic and using the boundary between a5 and a6 as a reference point. c1, c2, c3, c4 ... that are calculated by sequentially shifting the window to the right by tic are called the series of the cohesion scores of window width w from the leading part of the document to the end.

[0063]

Fig. 13 shows a graph prepared in such a way that the total number of the content words that appear between the beginning of a document and each reference point is set as a horizontal axis, and the cohesion score series of the minimum window width (40 words) is plotted in the above-mentioned word recognition result. For example, in the case of the cohesion score c2 of Fig. 12, the total number of the content words in the areas a1 through a5 is indicated by the position of the reference point. Here, the cohesion score is calculated from the beginning of the document toward the end by using 1/8 (5 words) of the window width of 40 words as an interval width tic.

[0064]

2. Recognition of the hierarchical relation of topics

The thematic boundary candidate sections that are obtained using the different window widths are integrated, and the hierarchical structure of topics and thematic boundaries are determined.

[0065]

Fig. 11 is a flowchart showing the thematic hierarchy recognition process performed by the thematic hierarchy detector 25. The thematic hierarchy detector 25 first receives three parameters of the biggest window width w1, the minimum window width w_min, a window width ratio r from a user (step S41), and obtains a set W of window widths to measure the cohesion score (step S42). The set W of window widths is generated by collecting terms equal to or more than w_min from the geometric series the first term of which is w1 and the common ratio of which is 1/r.

[0066]

At this time, it is sufficient in practical use to give about 1/2 to 1/4 of the size of the entire document for the biggest window width w1 in W, to give the words equivalent to a paragraph (for example 40 words) for the minimum window width w_min, and to give 2 for the window width ratio r.

[0067]

Next, the thematic boundary candidate section recognition process performed in step S44 of Fig. 11 is described using Figs. 12 and 14. The moving average method that is used here is a statistical method for the time series analysis to obtain the trend of a general situation by removing the small variation, which, for example, is used for the analysis of the variation of a stock price, etc. In the present embodiment, the moving average value of the cohesion score series is not only used to disregard the small variation, but also is considered as a forward cohesion force at the starting point of the moving average and as a backward cohesion force at the end point of the moving average. In this way, the value is a direct clue for the thematic boundary candidate section recognition.

[0068]

Fig. 12 shows the relationship between the cohesion score series of c1 through c4 and document areas a1 through a11 as mentioned above. The moving average value of the cohesion score series is an arithmetic mean score of continuous n pieces in the cohesion score series, such as, (c1+c2)/2 (two-item moving average), (c1+c2+c3)/3 (three-item moving average), (c1+c2+c3+c4)/4 (four-item moving average).

[0069]

Fig. 14 shows the contribution of a document area to the moving average of the cohesion score series shown in Fig. 12. Here, three kinds of the moving average of two to four terms of the cohesion scores of Fig. 12 are shown as an example. The figures in the table indicate the number of times that each document area is used when the cohesion score related to a moving average is calculated. Of these values, an underlined value indicates that the corresponding document area is used for the calculation of all the cohesion scores that are related to the moving average.

[0070]

For example , a value "1" on the upper-left corner shows that the document area a1 is included in a part of the left window only once in the moving average of four terms c1 through c4. Also, a value "2" on the right of the corner shows that the document area a2 is included in a part of the left window twice in the moving average of four terms c1 through c4. Regarding other numbers times in use, the meaning is the same.

[0071]

Since a cohesion score is an index for indicating the strength of a relation between parts before and after the boundary, a moving average value calculated using a cohesion score c1 that is obtained by including the area a1 in the left window, also indicates whether the area a1 is related rightward.

[0072]

In other words, it can be said that the moving average value indicates the strength of forward cohesion (forward cohesion force), i.e., how strongly the areas in the left window part with which the moving average value is calculated (a1 through a7 areas for the average of four terms c1 through c4) are pulled in the direction to the end of the document (forward direction: right direction in Fig. 15). On the other hand, it can be also said that the moving average value indicates the strength of backward cohesion (backward cohesion force), i.e., how strongly the areas in the right window part with which the moving average value is calculated (a5 through a11 areas for the average of four terms c1 through c4) are pulled in the direction of the leading part of the document (backward direction: left direction in Fig. 15).

[0073]

When studying the relevance between the cohesion force and each document area, it is conceivable that the more times an area is included in the window when a cohesion force is calculated, the stronger the contribution of that area to that force is. Since it is generally conceivable that the lexical cohesion is strong when the vocabularies are repeated in the vicinity, the contribution of the area that is close to the reference point (boundary position between the right window and the left window) of the cohesion score is strong.

[0074]

For example, regarding the moving average of four terms of Fig. 14, four boundaries between a4 and a5, a5 and a6, a6 and a7, and a7 and a8 are set as the reference points of the cohesion score. In this case, it is understood that a4 is included in the left window most frequently, and is the closest to the reference point. Also, it is understood that a8 is included in the right window most frequently, and is the closest to the reference point. Therefore, the area having the strongest relationship with the moving average value is a4 for the left window and a8 for the right window.

When similarly choosing the area having the strongest relationship with the moving average of three terms, a4 is obtained for the left window and a7 is obtained for the right window. Further when choosing the area having the strongest relationship with the moving average of two terms, a4 is obtained for the left window and a6 is obtained for the right window. The number of times in use of these areas is shown being enclosed with the frame of a thick line in Fig. 14.

[0076]

On the basis of the above-mentioned study, the thematic boundary detector 26 handles the moving average value of the cohesion score both as the index of the forward cohesion force at the first reference point inside the area for which the moving average is calculated and as that of the backward cohesion force at the last reference point. For example, the moving average value of four terms c1 through c4 becomes the forward cohesion force at the boundary of a4 and a5 and the backward cohesion force at the boundary of a7 and a8.

[0077]

Fig. 15 is a flowchart of the thematic boundary candidate section recognition process performed by the thematic boundary detector 26. The detector 26 first receives the interval width tic of the cohesion score series and the number n of terms of the moving average from the user (step S51).

[0078]

As for the rough standards of the values of these parameters, the interval width tic is about 1/8 to 1/10 of the window width w, and the number n of terms is about the half of w/tic (4 to 5). Further, the distance from the first to the last reference points of the area for which the moving average is calculated is computed by (n-1)*tic, and the computed value is made the width (word) of the moving average.

[0079]

Then, the moving average of the cohesion score is computed within the range of p to p+d for each position p in the document, and the average value is recorded as the forward cohesion force at the position p (step S52). This value is simultaneously recorded as the backward cohesion force at the end position p+d of the range within which the moving average is computed.

[0080]

Then, the difference between the forward cohesion force and backward cohesion force in each position (forward cohesion force -backward cohesion force) is checked from the beginning of the document toward the end, based on the recorded forward cohesion force. The position where the difference changes from negative to positive is recorded as a negative cohesion force equilibrium point mp (step S53).

[0081]

The negative cohesion force equilibrium point is a point such that the backward cohesion force is superior at the left of the point, and that the forward cohesion force is superior at the right of the point. Therefore, it is conceivable that the connection of the left and right parts is weak. Therefore, the negative cohesion force equilibrium point becomes the position candidate of a topic boundary.

[0082]

Then, a range [mp-d, mp+d] within d words immediately before and immediately after the recorded negative cohesion force equilibrium point mp is recognized as the thematic boundary candidate section (step S53), and the process terminates.

[0083]

Here, the meaning of recognizing the thematic boundary candidate section on the basis of the difference between the forward cohesion force and the backward cohesion force is described using Fig. 16. Fig. 16 shows the distribution of the cohesion score of the forward and backward cohesion forces using the window of 40-word width in the vicinity of 400 words (370 to 400 words) of Fig. 13. As for the interval width tic, 1/8 of the window width is adopted.

[0084]

In Fig. 16, the polygonal line graph plotted by a symbol '+' shows the series of cohesion force C, the polygonal line graph plotted by a symbol '*' shows the series of forward cohesion force FC, and the polygonal line graph plotted by a symbol '□' shows the series of backward cohesion force BC. The area shown by a rectangle indicating the thematic boundary candidate section will be described later.

[0085]

Furthermore, ep1, ep2, and ep3 that are shown by the dotted lines indicate three points (cohesion force equilibrium points) where the difference between the forward cohesion force and the backward cohesion force becomes 0. At the left side of the first point ep1, the backward cohesion force is superior to the forward cohesion force. From the right side of ep1 to the next point ep2, the forward cohesion force is superior to the backward cohesion force. Furthermore, from the right side of pl2 to the last point ep3, the backward cohesion force is superior to the forward cohesion force. At the right side of ep3, the forward cohesion force is superior to the backward cohesion force.

[0086]

Therefore, ep1 and ep3 are the negative cohesion force equilibrium points where the difference between the forward cohesion force and the backward cohesion force changes from negative to positive, and ep2 is the positive cohesion force equilibrium point where the difference changes from positive to negative.

[0087]

It is understood from such a change of cohesion force that the area on the left side of the point ep1 shows the comparatively strong cohesion with any part on the further left side, the areas of both sides of the point ep2 in the middle show the strong cohesion toward ep2, and the area on the right side of the point ep3 shows comparatively strong cohesion with any part on the further right side. Actually, the cohesion score that is plotted with the forward and backward cohesion forces takes a minimal value at the vicinity of ep1 and ep3, and takes the maximal value at the vicinity of ep2. In this way, the change of the forward and backward cohesion forces is closely related to the change of cohesion score.

[0088]

For example, there is a minimal point (in this case, c3) of cohesion score series in the vicinity of the cohesion force equilibrium point ep3 of Fig. 16. The minimal value of FC and BC showed by an upward arrow is the value that is obtained by moving-averaging the cohesion scores (four terms of c1 to c4) of a horizon arrow part. In this way, the cohesion force usually takes a minimal value in the vicinity (within the width of the moving average) corresponding to the minimal point of the cohesion score. If there is a small variation in a narrower range than the area where the moving average is computed, however, sometimes the moving average value (i.e., cohesion force) does not take a minimal value due to the smoothing function of the moving average.

[0089]

Since the forward cohesion force is moving average value recorded at the starting point of the area where the moving average is computed, the minimal position of the forward cohesion force is at the left of the minimal position of the cohesion score. Similarly, the minimal position of the backward cohesion force is at the right of the minimum position of a cohesion score. Then, a cohesion force equilibrium point is formed in the area where the moving average is computed if the variation of the cohesion score is sufficiently large.

[0090]

Fig. 17 is a flowchart showing the thematic boundary recognition process that is carried out in step S45 of Fig. 11. The thematic hierarchy detector 25 sorts the recognized thematic boundary candidate sections using the window width of the cohesion score series and the appearance position in the document of the cohesion force equilibrium point of the thematic boundary candidate section, and generates the series B(j)[p] of thematic boundary candidate section (step S61).

[0091]

Here, a control variable j is the series number showing that the cohesion score series were calculated by window width wj. A control variable p is the data number for each thematic boundary candidate section of the series. In reality, the control variable j takes 1, 2,... in order from the largest window width. The control variable p takes 1, 2,... in the appearance order of the cohesion force equilibrium point. Each data B(j)[p] includes the following element data.

[0092]

B(j)[p].range: Thematic boundary candidate section. (a starting position and an end position)

B(j)[p].ep: Cohesion force equilibrium point.

[0093]

B(j)[p].child: Thematic boundary candidate section (child candidate section) of B(j+1) series that coincides with the thematic boundary candidate section of the boundary position in its range.

A cohesion force equilibrium point is a point theoretically. However, since the point where the sign of the difference between the forward cohesion force and backward cohesion force reverses is recognized as the equilibrium point as mentioned above, the point is actually expressed by a set of the negative point (starting position) and the positive point (end position). Thereupon, in this preferred embodiment, the values (forward cohesion force-backward cohesion force) at the starting position lp and the end position rp of the cohesion force equilibrium point are set as DC(lp) and DC(rp), respectively, and a point ep where the cohesion force at the right and left becomes 0 is obtained by interpolating the following equation:

$$ep=(DC(rp)*lp-DC(lp)*rp)/(DC(rp)-DC(lp)) \qquad (2)$$

Then, the obtained ep is set as B(j)[p].ep.

[0094]

Then, the thematic hierarchy detector 25 correlates the thematic boundary candidate section data having different window widths. Here, a plurality of pieces of B(j)[p] that belong to one series are grouped to be described as B(J), and furthermore, the following processes are described using the following notation.

[0095]

ie: Series number corresponding to the minimum window width w_min.

|B(j)|: Maximum value of data number p in B(j).

First, series number i indicating the data to be processed is initialized to 1 (step S62). In this way, the series of the thematic boundary candidate section obtained by the maximum window width w1 is set as the data to be processed. As long as j+1≤je, a correlation process of setting B(j+1) as the series to be related to is performed while incrementing j.

[0096]

In this correlation process, for each thematic boundary candidate section datum B(j)[p] (p=1, ..., |B(j)|) in the series to be processed, the datum of which B(j+1)[q].ep is the closest to B(j)[p].ep is chosen among data B(j+1)[q] of the series to be correlated. The chosen datum is stored in B(j)[p].child as correlated boundary candidate section data.

[0097]

The concrete procedures are as follows: first, j+1 and je are compared (step S63). If j+1≤je, 1 is assigned to p (step S64), and p is compared with |B(j)| (step

S65). If $p \leq |B(j)|$, correlation processes in and after step S66 are performed. If p exceeds $|B(j)|$, j=j+1 is set (step S71), and the processes in and after step S63 are repeated.

[0098]

In step S66, the thematic hierarchy detector 25 selects the data B(j+1)[q] in which B(j+1)[q].ep∈B(j)[p].range and in which B(j+1)[q].ep is the closest to B(j)[p].ep from the candidate data B(j+1)[q] (q=1,..., $|B(j+1)|$) as the data to be correlated. Then, the selected data is stored in B(j)[p}.child.

[0099]

Here, the condition of B(j+1)[q].ep∈B(j)[p].range shows that the cohesion force equilibrium point of B(j+1)[q] is included in the thematic boundary candidate section of B(j)[p].

[0100]

Fig. 18 shows a selection example of data to be correlated. In Fig. 18, the polygonal line graph plotted by the symbol '+' shows the series of the forward cohesion force by the window of 80-word width corresponding to the data to be processed. The polygonal line graph plotted with the symbol '×' shows the series of the backward cohesion force by the window of 80-word width. The polygonal line graph plotted with the symbol '*' shows the series of the forward cohesion force by the window of 40-word width corresponding to the data to be processed. The polygonal line graph plotted with the symbol '□' shows the series of the backward cohesion force by the window of 40-word width. Furthermore, the area shown by a rectangle corresponds to the thematic boundary candidate section, and ep1 and ep3 that are shown by the dotted line correspond to the equilibrium point of the cohesion force by the window of 40-word width.

[0101]

For example, when the datum to be processed is set as B(3)[4], there are cohesion force equilibrium points of ep1 and ep3 in its vicinity, and there are two pieces of data B(4)[6] and B(4)[7] of the series to be correlated in correlation with the data to be processed. Since of these, the cohesion force equilibrium point ep3 of B(4)[7] is included in the thematic boundary candidate section (rectangle in the upper section) of B(3)[4], B(4)[7] is selected as a datum to be correlated.

[0102]

Then, the thematic hierarchy detector 25 determines whether the datum to be correlated is selected (step S67). If the datum to be correlated is selected, p=p+1 is set (step S70), and the processes in and after step S65 are repeated.

[0103]

If the datum to be correlated which meets the condition is not detected, a dummy datum B(j+1)[q] which has the same thematic boundary candidate section as B(j)[p] is generated to be inserted into the series of B(j+1) (step S68).

[0104]

In step S68, first the value of B(j)[p].range and B(j)[p].ep are set to B(j+1)[q].range and B(j+1)[q].ep, respectively, and a new datum B(j+1)[q] is generated. The generated datum B(j+1)[q] is inserted into a position where B(j+1)[q-1].ep<B(j+1)[q].ep and B(j+1)[q].ep<B(j+1)[q+1].ep in the series B(j+1).

[0105]

In this way, a data number q of the dummy datum is decided, and the data number of the subsequent existing data is rewritten. It is because in subsequent processes, a topic sentence is extracted in the thematic hierarchies of all series numbers after j to generate the dummy thematic boundary candidate section datum.

[0106]

Then, the generated dummy datum B(j+1)[q] is stored in B(j)[p].child (step S69), and the processes in and after step S70 are performed. If j+1 exceeds je in step S63, the processes terminate.

[0107]

Finally, for each datum B(j)[p] of all the series number j that is smaller than je, the data of series number j+1 that has the cohesion force equilibrium point in the thematic boundary candidate section B(j)[p].range is set in B(j)[p].child. Therefore, the thematic boundary candidate section data of a plurality of hierarchies is correlated to each other in a chain by B(j)[p].child.

[0108]

Fig. 19 shows the recognition result of the thus-obtained thematic boundary. In Fig. 19, the bar chart in which window widths (vertical ordinate axis) are 320 words, 160 words, 80 words and 40 words shows the final thematic boundary of the topic of grading corresponding to each window width, specifically, the position of the cohesion force equilibrium point of the minimum window width (40 words). The rectangular area that intersects the bar chart shows the thematic boundary candidate section that is recognized by the cohesion force of each window width.

[0109]

In step S46 of Fig. 11, the thematic boundary shown in Fig. 19 is finely adjusted to be matched with the starting position of the sentence, thereby generating a thematic hierarchy such that one topic is set between the boundaries. Thus, part of the thematic boundary of Fig. 19 is shifted by this fine adjustment, and consequently the thematic hierarchy of the tree-structure as shown in Fig. 20 is generated.

[0110]

For example, by the boundary corresponding to the minimum window width, 15 topics shown in Fig. 20 are recognized as the lowest hierarchical topics in correlation with 15 sections marked with an arrow. By the boundary according to a window width on 80 words, of the 15 topics, ten topics in total that are obtained by incorporating topics corresponding to four groups of sections 2 and 3, sections 4 through 6, sections 11 and 12 and sections 13 and 14 are recognized as the second hierarchical topics.

[0111]

In the thematic hierarchy of Fig. 20, the node that is shown by a rectangle corresponds to each recognized topic, and the number inside the rectangle corresponds to the section number shown in Fig. 19. Furthermore, the thematic hierarchy shown in Fig. 21 is generated by applying the same process to the second document to be read.

[0112]

Next, the process of the topic extractor 27 is described. Fig. 22 is a flowchart showing the topic extraction process performed by the topic extractor 27. The topic extractor 27 first receives two thematic hierarchies T1 and T2 of the first and the second documents to be read, respectively (step S101). Then, the extractor 27 calculates the relevance scores of all the topic sets (t1, t2) of a topic t1 in the thematic hierarchy T1 and a topic t2 in the thematic hierarchy T2 (step S102).

[0113]

In this preferred embodiment, the relevance score R(t1, t2) between the topics t1 and t2 is obtained by the similarity of the vocabulary that is included in the sections s1 and s2 of the document, corresponding to t1 and t2, respectively. Specifically, R(t1, t2) is calculated by the following equation.

[0114]

[Mathematics 2]

$$R(t1, t2) \equiv R(s1, s2) = \frac{\sum_t w_{t,\,s1} w_{t,\,s2}}{\sqrt{\sum_t w^2_{t,\,s1} \sum_t w^2_{t,\,s2}}} \qquad (3)$$

Here, $w_{t,s1}$, $w_{t,s2}$ respectively represent the weight that indicates the importance of a word t in sections s1 and s2, and is calculated by the following equation.

[0116]

[Mathematics 3]

$$w_{t,s} = tf_{t,s} \times \log\left(\frac{|D|}{df_t}\right) \qquad (4)$$

[0117]

In equation (4), $tf_{t,s}$ represents the appearance frequency of word t in section s. $|D|$ represents the number of blocks that are obtained by dividing the document including section s for each fixed word width (80 words), and df_t represents the number of blocks where word t appears.

[0118]

The equations (3) and (4) are variations of a calculation method called tf·idf method to be used for the calculation of a query-document relevance score in the information retrieval field. According to the tf·idf method, the part $|D|/df_t$ of the equation (4) is calculated in the units of the number of documents that are included in the document aggregate to be retrieved not in the units of the number of sections of a document. Specifically, if $|D|$ is set as the number of documents in the document aggregate to be retrieved and df_t is set as the number of documents where the word t appears, these equations become equivalent to a general tf·idf method calculation equation.

[0119]

The relevance score R(t1, t2) may be obtained using the tf·idf method. However, since the relevance score can be calculated only from the document to be read according to the equations (3) and (4) of this preferred embodiment, and also a sufficiently effective result can be obtained by these calculation equations, as described later, these calculation equations are selected here.

[0120]

Then, the topic extractor 27 calculates threshold values used to select a topic pair from all the combinations of topics t1 and t2 of the first and the second documents to be read utilizing a thematic hierarchy. As the threshold, for example, the maximum relevance score of the sub-tree of the thematic hierarchy is used. Here, the maximum relevance score in the sub-tree below a certain topic t is the maximum value of the relevance score that is calculated for t or the descendant of t (any smaller topic that constitutes t) in the thematic hierarchy.

[0121]

The topic extractor 27 first obtains the maximum relevance score for topic t1 and records it on t1.max (step S103). It then similarly records the maximum relevance score topic t2 on t2.max (step S104). Then, the extractor 27 obtains an aggregate of topic pairs T that are defined by T≡{(t1, t2) | R(t1, t2)≥max(t1.max, t2.max), outputs it as the common topics (step S105), and terminates the process.

[0122]

A specific example of the topic extraction process based on the maximum relevance score is described using Figs. 23 and 24. Fig. 23 shows the calculation result of relevance scores in step S102 of Fig. 22. Here, all topic pairs with a relevance score 0.25 or more are shown by the dotted arc. The numerical value attached to the arc indicates a relevance score. The left graph of the two tree-structure graphs corresponds to the thematic hierarchy of Fig. 20 while the right graph corresponds to the thematic hierarchy of Fig. 21.

[0123]

Note a node in lower right corner of the right graph (hereinafter called "node right 7"). This node a node indicating a topic corresponding to the last minimum section of the second document to be read and is a leaf node (node having no child) in the graph.

[0124]

Therefore, the maximum relevance score of this node is the maximum relevance score of an arc that is directly connected to this node. In node right 7, the relevance score 3.0 of the topic pair of node left 13-14, node right 7) is the maximum relevance score. Since there is no arc having a relevance score exceeding 0.35 in node left 13-14, the topic pair of (node left 13-14, node right 7) is output as a common topic.

[0125]

As for node right 6-7, since the sub-tree below this node includes node right 7, the topic pair of an arc directly connected to node right 6-7 is not output as a common topic unless its relevance score at least exceeds the maximum relevance score (0.35) of node right 7. Since node right 6-7 has no such arc, the topic pair including node 6-7 is output as a common topic.

[0126]

In this way, by selecting the topic pairs using the maximum relevance score in the sub-tree, common topics between the two documents to be read can be narrowed down to the topic pair shown in Fig. 24. In Fig. 24, although only seven pairs of common topics other than the ones that correspond to the relation between the entire documents are extracted, the topics that do not belong to any topic pairs are only three of node left 1, node left 11 and node left 15. Of these topics, the node that directly describes no question item is only node left 15, and the remaining is the content part not directly connected to a question/answer part that plays a role of introducing a subsequent topic.

[0127]

Any of the extracted seven pairs of common topics includes appropriately corresponding contents, as shown by a result described later. In this way, according to this preferred embodiment, an appropriate pair of topics can be selected neither excessively nor insufficiently without setting a special threshold in advance, by selecting the common topics utilizing the thematic hierarchies.

[0128]

Then, for each topic pair that is extracted by the topic extractor 27, the output unit 28 extracts a related part corresponding to the topic pair from each document to be read and outputs the related part. For example, regarding the topic pair of relevance score 0.30 of (node left 9-10, node right 4-5) of Figure 24, sections 9 and 10 in the first document to be read are extracted corresponding to the topic of node left 9-10, and sections 4 and 5 in the second document to be read are extracted corresponding to the topic of node right 4-5. Then, the sections are arranged in such a way that a user can easily contrast them and the thus-rearranged sections are output.

[0129]

Fig. 25 shows an example of the output result of the related parts for this topic pair. In the output example of Fig. 25, the left column shows a related part of the first document and the right column shows a related of the second document. Each related part is divided into the units of minimum topics (minimum section) that are recognized by the thematic hierarchy detector 25. The words emphasized with a bold font are those words that appear in both the columns and have relatively high importance calculated by the equation (4) in each related part. Specifically, these words are extracted with the following procedures.

[0130]

The words that appear in both related parts are first extracted as keyword candidates. For each extracted word, the value of the equation (4) in each related part is obtained as the importance of each word. Then, keywords are extracted in order of importance until the accumulated value of the importance of the extracted words exceeds 1/2 of the total value of the importance of the whole candidates.

[0131]

The common topic shown in Fig. 25 is only a part extracted duplicated in the document pair to be read used in this preferred embodiment. In this part, not only a topic pair in an upper layer (node left 9-10, node right 4-5) but also two topic pairs (node left 9, node right 4) and (node left 10, node right 5) are extracted as common topics.

[0132]

As seen in the contents shown in Fig. 25, it is conceivable that the answer to the question of node left 9 is node right 4, and the answer to the question of node left 10 is node right 5. The strong relevance can be recognized between the node left 9 and node 10 left, and the node right 4 and node right 5.

[0133]

Therefore, although these related parts are duplicated and extracted, it is understood that the relationship between aggregates of two nodes each and the relationship between individual nodes are not redundant but express important correspondence. Thereupon, in Fig. 25, the starting positions of the corresponding topic are aligned to be output so that not only the two entire aggregates of topics but also the individual topics can be contrasted.

[0134]

Furthermore, the output unit 28 can also improve the taking-a-look efficiency by summarizing and displaying the contents of the related part. If, the technology disclosed in, for example, above-mentioned Japanese Patent Laid-open Publication No. 11-272,699 is used, a concise summary that includes a lot of keywords extracted in the above-mentioned procedures can be generated.

[0135]

Fig. 26 is a flowchart showing the simplified procedures of such a summarizing process. The output unit 28 first receives a related part P1 that is extracted from the first document and a related part P2 that is extracted from the second document in correlation with to a common topic (step S121). Then, the output unit 28 extracts keywords from each of the related parts P1 and P2, and merges those keywords (step S122).

[0136]

Then, the output unit 28 selects important sentences from the related part P1 and generates a summary (step S123), and similarly generates a summary from the related part P2 (step S124). Then, the unit 28 arranges the summaries so as to be easily compared, and outputs the summaries side by side (step S125), thereby terminating the processes.

[0137]

Fig. 27 is a flowchart showing the important sentence selecting process performed in steps S123 and S124 of Fig. 26. In this process, the output unit 28 first sets P1 or P2 as target parts P, and stores the keywords extracted in step S122 in an keyword list KWL as the clues of an important sentence (step S131). Then, the output unit 28 selects a sentence that includes the most number of keywords from the target parts P (step S132), and determines whether such a sentence can be selected (step S133).

[0138]

If such a sentence can be selected, the keywords included in the selected sentence are removed from the KWL (step S134), and determines whether the KWL is empty (step S135). If the KWL is not empty, the processes in and after step S132 are repeated. Then, the processes terminate when at least one important sentence can be selected for all the keywords. The output unit 28 arranges the selected sentences in order of appearance in the original document, and outputs the sentences as summaries (step S136), thereby terminating the processes.

[0139]

If in step S133 no sentence including a keyword can be selected at all, the process is terminated and the process in step S136 is performed. By performing the processes shown in Figs. 26 and 27, summaries shown in Figs. 28, 29, and 30 are generated.

[0140]

In this way, not only by separately presenting the related parts corresponding to each common topic, but also by summarizing the related parts, a list of related parts can be output in such a way that a user can easily take a look. Therefore, even if many common topics are extracted at once when comparing/reading a long document, the output unit 28 can effectively support the comparison/reading work.

[0141]

Furthermore, the output unit 28 can support the work of analyzing the related parts while studying the positioning the related parts in the entire documents to be read, by displaying related parts the documents to be read, which is the original documents, side by side. In this case, it is sufficient to display the summaries of the related parts and the entire documents to be read side by side, as shown in Fig. 31. Furthermore, the reading efficiency can be enhanced more, if a hyper-link is provided between a related part and the corresponding part of the entire document to be read.

[0142]

In Fig. 31, the left frame is the window for the reference of related parts. The right frame is the window for the reference of the documents to be read. In the left frame, the generated summaries of the extracted parts are displayed, and the anchor of the hyper-link for the target part of the document to be read is set in the key-parentheses (underlined part) after the speaker's name. By a user designating the anchor as occasion demands, the designated part of the first document to be read is displayed on the upper right window, and the designated part of the second document to be read is displayed on the lower right window.

[0143]

In the document to be presented in the right frame, the related parts are highlighted by an underline, so that the related parts can be distinguished from the context before or after them. As for the method of highlighting display, color display, half-tone dot meshing, etc. can be used. In this example, the summaries of the extracted parts are displayed in the left frame. Instead, the extracted parts themselves may be displayed. Furthermore, it is conceivable that the output unit 28 can switch between the presentation of the summary of the related part and the presentation of the entire contents of the related part, according to the request from a user.

[0144]

The output unit 28 can also display the relationship between the related parts of the documents using a graph, so that a user can understand the holistic relevance between the documents to be read with a glance. Fig. 32 shows an example of that presentation of the appearance state of the related parts in the documents to be read.

[0145]

In Fig. 32, the thematic hierarchies of two documents to be read are displayed at the top of the frame in a graph similar to that shown in Fig. 24. At the bottom of the frame, the first and second documents to be read are displayed side by side. To the graph of the thematic hierarchies, arcs that indicate the common topic are added, and each arc is provided with a hyper-link for synchronizing the related parts of both the documents to be read. Furthermore, at each node corresponding to the topic, a hyper-link for the corresponding part of each document to be read is provided. The related part of each document to be read is highlighted similarly to that of Fig. 31.

[0146]

In this way, the understanding of the holistic relevance between the documents to be read is promoted by presenting the appearance state of the related parts between the documents to be read.

Thus, a user can understand with a glance whether the documents to be read corresponds to each other one to one as a whole like the documents to be read in this preferred embodiment or in the documents to be read related parts collectively appear in a specific part. If the latter is the case, the user can select the part in which the related parts are collected and efficiently read the documents.

[0147]

In the above-mentioned preferred embodiment, two documents to be read are mainly described, but the comparison and reading of three or more documents can also be supported by applying this process. In this case, the above-mentioned process can be also performed, for example, by setting any one of the documents as the reference (axis) and comparing other documents, or by performing the process like the above-mentioned to all the combinations of the documents to be read, and then by arranging and integrating the extracted topics by any means. If the latter is the case, it is conceivable that the topics of the other documents corresponding to the same part of any one document are integrated.

[0148]

For example, when the total eight representative questions (excluding answers) included in the above-mentioned " the minutes No. 2 of the 149th plenary session of the House of Representatives" are extracted as different documents and are compared using the general policy speech of the Prime minister in "The minutes No.1 of the 149th plenary session of the House of Representatives" (on July 28, 2000) as a reference document, the related part as shown in Fig. 33 is extracted as the related part that is also related to the above-mentioned first document to be read (representative question by Diet member Mizushima).

[0149]

In Fig. 33, the left column corresponds to the summary of the related part of the reference document, the central column corresponds to the summary of that of the first document, and the right column corresponds to that of the other document. Here, although only the part related to the first document to be read is shown as an example, similarly the representative question made by the other questioner can be also corresponded to the appropriate part of the reference document.

[0150]

Furthermore, such a related part can be combined with the reference document to be output. In this way, the generation of an integrated document such as "the point of the policy speech and the view of each party representative to the speech" can be supported.

[0151]

Fig. 34 is a flowchart showing such a document integration process. The document reading apparatus firstly selects a reference document from a plurality of documents to be read on the basis of instructions, etc. from a user (step S141), and extracts the related parts of the other document related to the reference document (step S142). Then, the output unit 28 merges the extracted related parts in order of appearance in the reference document, generates an integration document (step S143), outputs the document (step S144), and terminate the process.

[0152]

Next, the process of English document is described exemplifying the case where two communiqués by G8 of the Kern summit in 1999 and the Okinawa summit in 2000 are targeted. Here, the first English document to be read and the second English document to be read are set as Foreign 1 and Foreign 2, respectively. [0153]

[Foreign 1]

"G8 COMMUNIQUÉ KÖLN 1999"

[Foreign 2]

"G8 COMMUNIQUÉ OKINAWA 2000"

[0155]

All the sentences of these two documents are composed of 4500 words and 7000 words individually. Since it is too long to describe all the processing results in the specification and drawings, only the half of them is targeted in the following process. In the first document to be read composed of ten paragraphs as a whole, the following five paragraphs (1800 words) are targeted to be processed, while in the second document to be read, the following one part (3500 words) that is located next to the preamble is targeted to be processed.

(1) Part to be processed of the first document to be read

I. Getting the World Economy on Track for Sustained Growth II. Building a World Trading System That Works for Everyone

III. Designing Policies for More Employment

IV. Investing in People

V. Strengthening Social Safeguards

(2) Part to be processed of the second document to be

read

Toward a 21st century of greater prosperity

Furthermore, the following processing method and parameters are adopted here.

(1) Method of word recognition: Method using a stop word list

(2) Width of the window for cohesion score calculation:

Minimum window width: w_min=80 (word)

Maximum window width w1: The number of words of the value that is equal to a product obtained by multiplying w_min with 2**n (n-th power of 2) and does not exceed the half of all the documents

Interval width: 1/8 of window width

Fig. 35 shows the leading part of the first document to be read. Fig. 36 shows the processing result of tokenizer 22 for the part. In Fig. 36, a part enclosed with [ ] corresponds to recognized word. The word only the leading letter of which is capitalized is rewritten into all the small letters in the [ ].

[0156]

In this case, the tokenizer 22 extracts words using a space and delimiter symbols such as ",", ".", ":", ";", etc. as clues, and removes the words that are included in the stop word list as shown in Fig. 37, thereby recognizing words. The stop word list is a list for defining in advance words such as articles, prepositions, etc., that are not to be extracted as keywords.

[0157]

Fig. 38 shows the extraction result of common topics for the above-mentioned document pair. In Fig. 38, the left tree-structured graph corresponds to the output of the thematic hierarchy detector 25 for the first English document to be read, that is, corresponds to the recognition result of the thematic hierarchy of the first English document to be read. The right tree-structured graph corresponds to the recognition result of the thematic hierarchy of the second English document to be read. Also, the arc between these tree-structured nodes shows the related topic pair that is extracted by the topic extractor 27.

[0158]

When the output unit 28 summarizes the thus-extracted common topics in the procedures of Figs. 26 and 27, the summaries shown in Figs. 39, 40 and 41 are obtained.

As described above, the present invention is applicable to English documents similarly to Japanese documents. Furthermore, the present invention can be applied to documents written in any language or in any form, and can obtain approximately the same result.

(Appendix 1) A document reading apparatus presenting a plurality of documents designated as reading documents by a user, comprising:

a thematic hierarchy recognition device recognizing a thematic hierarchy of each of the plurality of documents;

a topic extracting device extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

a topic relation presentation device taking out a description part corresponding to the extracted topic from each of the plurality of documents and outputting the taken-out description parts.

(Appendix 2) The document reading apparatus according to claim 1, wherein regarding a topic set that consists of topics of various grading in the recognized thematic hierarchies, the topic extracting device calculates a relevance score between topics of the topic set based on lexical similarity of description parts corresponding to each topic of the topic set, and extracts a topic set having a relevance score equal to or more than a threshold that is set based on inclusive relationship of topics.

(Appendix 3) The document reading apparatus according to claim 1, wherein the topic relation presentation device presents the taken-out description parts side by side.

(Appendix 4) The document reading apparatus according to claim 3, wherein the topic relation presentation device presents the related parts and original documents in two windows, one of the windows including the related parts side by side and the other including the original documents side by side.

(Appendix 5) The document reading apparatus according to claim 3, wherein the topic relation presentation device presents summaries of the related parts.

(Appendix 6) The document reading apparatus according to claim 5, wherein the topic relation presentation device presents summaries of the related parts and original documents in two windows, one of the windows including the summaries side by side and the other including the original documents side by side.

The document reading apparatus according to claim 3, wherein the topic relation presentation device presents a plurality of thematic hierarchies corresponding to the plurality of documents and a correspondence relationship between the plurality of thematic hierarchies based on the plurality of common topics in a drawing, and presents a designated part of the plurality of documents in accordance with an instruction from the user given on the drawing.

(Appendix 7) The document reading apparatus according to claim 3, wherein the topic relation presentation device presents a plurality of thematic hierarchies corresponding to the plurality of documents and a correspondence relationship between the plurality of thematic hierarchies based on the plurality of common topics in a drawing, and presents a designated part of the plurality of documents in accordance with an instruction from the user given on the drawing.

(Appendix 8) The document reading apparatus according to claim 1, wherein the topic relation presentation device sets one document among the plurality of documents as a reference document, generates a new integrated document by merging the contents of the reference document with description parts of another document related to the reference document, and outputs the integrated document.

(Appendix 9) A computer-readable storage medium storing a program for a computer that presents a plurality of documents designated as reading documents by a user, the program causes the computer to perform:

recognizing a thematic hierarchy of each of the plurality of documents;

extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

extracting a description part corresponding to the extracted topic from each of the plurality documents and outputting the extracted description part.

(Appendix 10) A carrier signal carrying a program to a computer that presents a plurality of documents designated as reading document by a user, the program causes the computer to perform:

recognizing a thematic hierarchy of each of the plurality of documents;

extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

extracting a description part corresponding to the extracted common topic from each of the plurality documents and outputting the extracted description parts.

(Appendix 11) A document presenting method of presenting a plurality of documents designated as reading documents by a user, comprising:

recognizing a thematic hierarchy of each of the plurality of documents;

extracting a topic that commonly appears in the plurality of documents based on the recognized thematic hierarchies; and

extracting a description part corresponding to the extracted topic from each of the plurality documents and outputting the taken-out description parts.

[0159]

[Effect of the Invention]

Since according to the present invention, the topics of various grading in a plurality of documents to be read can be compared using the thematic hierarchy of each document to be read, the common topic the description amount of which largely differs from document to document can be extracted appropriately. Also, the related parts can be easily analyzed and compared by extracting the related part corresponding to the extracted common topic from each document to be read, and outputting the related parts side by side. Thus, the present invention can effectively support the comparative reading work of a plurality of documents.

[Brief Description of the Drawings]

Fig. 1 is a block diagram showing the principle of the document reading apparatus of the present invention.

Fig. 2 shows the basic configuration of the document reading apparatus of the present invention.

Fig. 3 shows the configuration of the information processor.

Fig. 4 shows the computer-readable storage medium.

Fig. 5 shows a document to be read.

Fig. 6 is a flowchart showing the word recognition process.

Fig. 7 shows an example of the word recognition result.

Fig. 8 is a flowchart showing the morphological analysis process.

Fig. 9 shows an example of how to consult a Japanese dictionary.

Fig. 10 shows an example of how to consult an English dictionary.

Fig. 11 is a flowchart showing the thematic hierarchy recognition process.

Fig. 12 shows an example of the series of the cohesion scores.

Fig. 13 shows an example of the cohesion score distribution.

Fig. 14 shows the contribution of a document area to the moving average.

Fig. 15 is a flowchart showing the thematic boundary candidate section recognizing process.

Fig. 16 shows an example of the cohesion force equilibrium point.

Fig. 17 is a flowchart showing the thematic boundary recognition process.

Fig. 18 shows an example of data to be correlated.

Fig. 19 shows an example of the thematic boundary recognition result.

Fig. 20 shows an example of the thematic hierarchy of the first document to be read.

Fig. 21 shows an example of the thematic hierarchy of the second document to be read.

Fig. 22 is a flowchart showing the common topic extraction process.

Fig. 23 shows an example of relevance score calculation result.

Fig. 24 shows an example of common topic extraction result.

Fig. 25 shows an example of the output related parts.

Fig. 26 is a flowchart showing the summarization process.

Fig. 27 is a flowchart showing the importance sentence selection process.

Fig. 28 shows an example of the summarized related parts (No. 1).

Fig. 29 shows an example of the summarized related parts (No. 2).

Fig. 30 shows an example of the summarized related parts (No. 3).

Fig. 31 shows an example of the presented related parts with an original document reference function.

Fig. 32 shows an example of the presented related parts with a graph.

Fig. 33 shows an example of the output related parts of three documents.

Fig. 34 is a flowchart showing the sentence integration process.

Fig. 35 shows an example of the leading part of an English document to be read.

Fig. 36 shows an example of the word recognition result of an English document to be read.

Fig. 37 shows an example of a stop word.

Fig. 38 shows the extraction result of an English common topic.

Fig. 39 shows an example of the summarized English related parts No. 1).

Fig. 40 shows an example of the summarized English related parts No. 2).

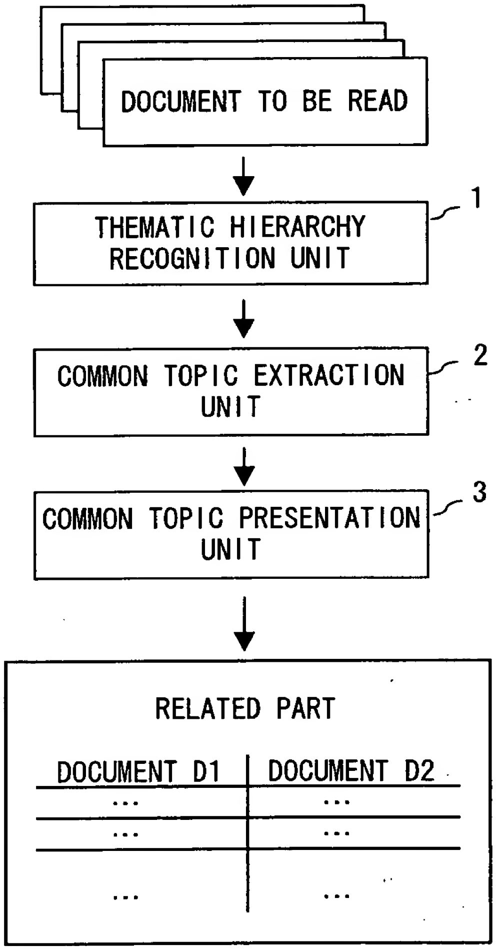
Fig. 41 shows an example of the summarized English related parts No. 3).

[Explanation of the Codes]

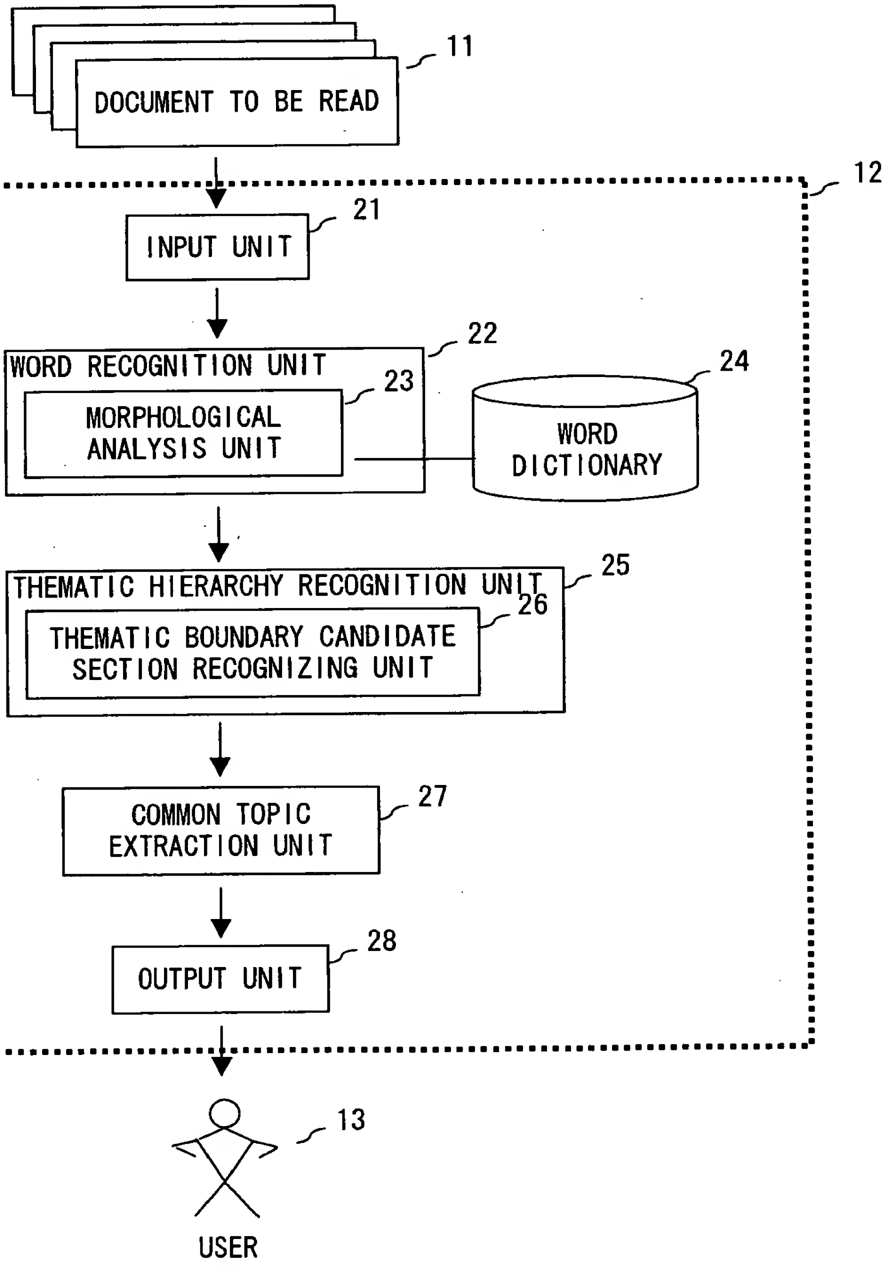
1	Thematic hierarchy recognition unit
2	Common topic extraction unit
3	Common topic presentation unit
11	Document to be read
12	Document reading apparatus
13	User
21	Input unit
22	Word recognition unit
23	Morphological analysis unit
24	Word dictionary
25	Thematic hierarchy recognition unit
26	Thematic boundary candidate section recognizing unit
27	Common topic extraction unit
28	Output unit
41	Output device
42	Input device
43	CPU
44	Network connection device
45	Medium driving device
46	Auxiliary storage device
47	Main storage
48	Bus

49	Portable storage medium
50	Server
51	Database

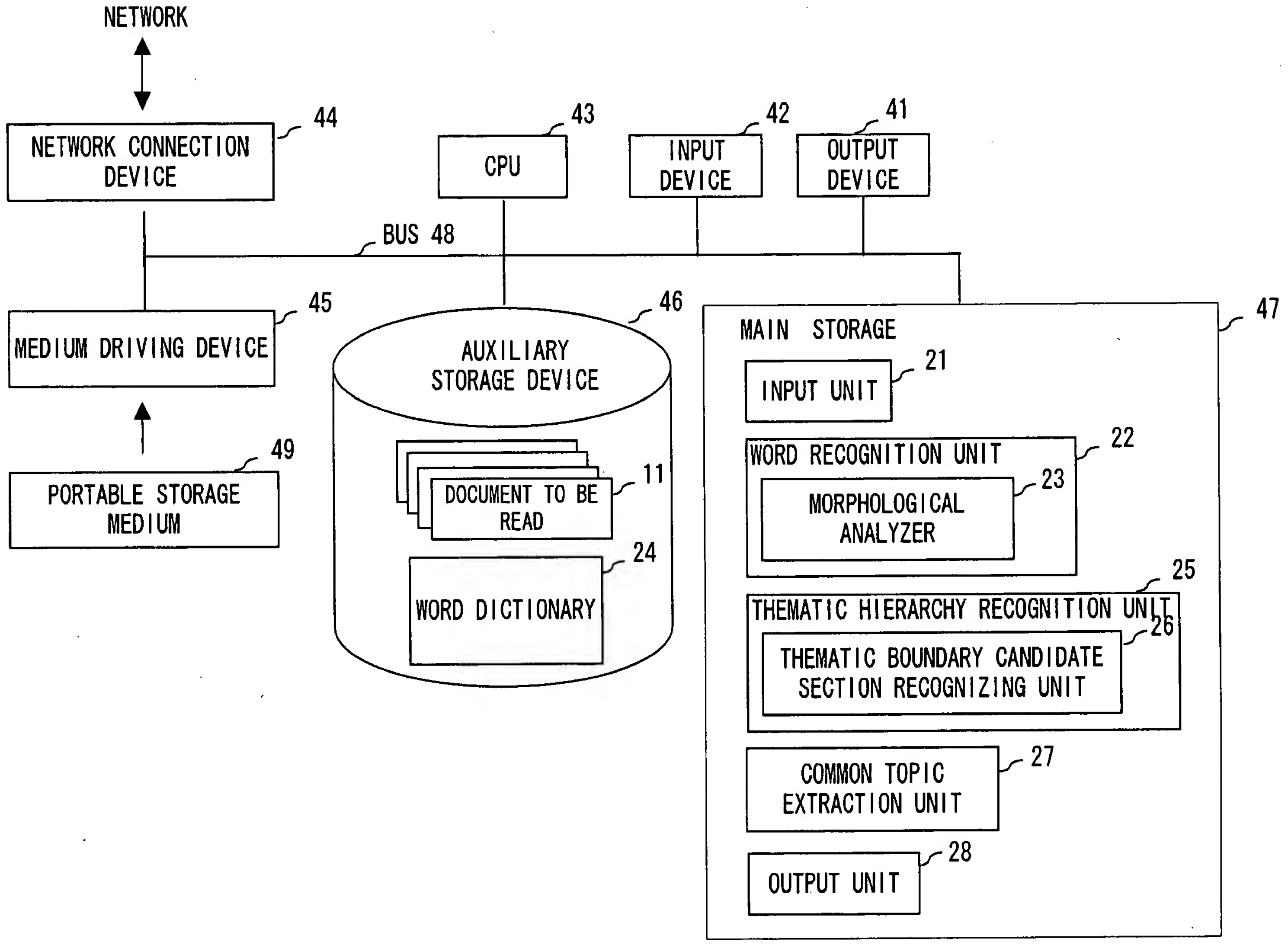
[FIG.1]
BLOCK DIAGRAM SHOWING THE PRINCIPLE OF THE PRESENT INVENTION


DOCUMENT TO BE READ
THEMATIC HIERARCHY RECOGNITION UNIT
1
COMMON TOPIC EXTRACTION UNIT
2
COMMON TOPIC PRESENTATION UNIT
3
RELATED PART
DOCUMENT D1
DOCUMENT D2
...
...
...
...
...
...

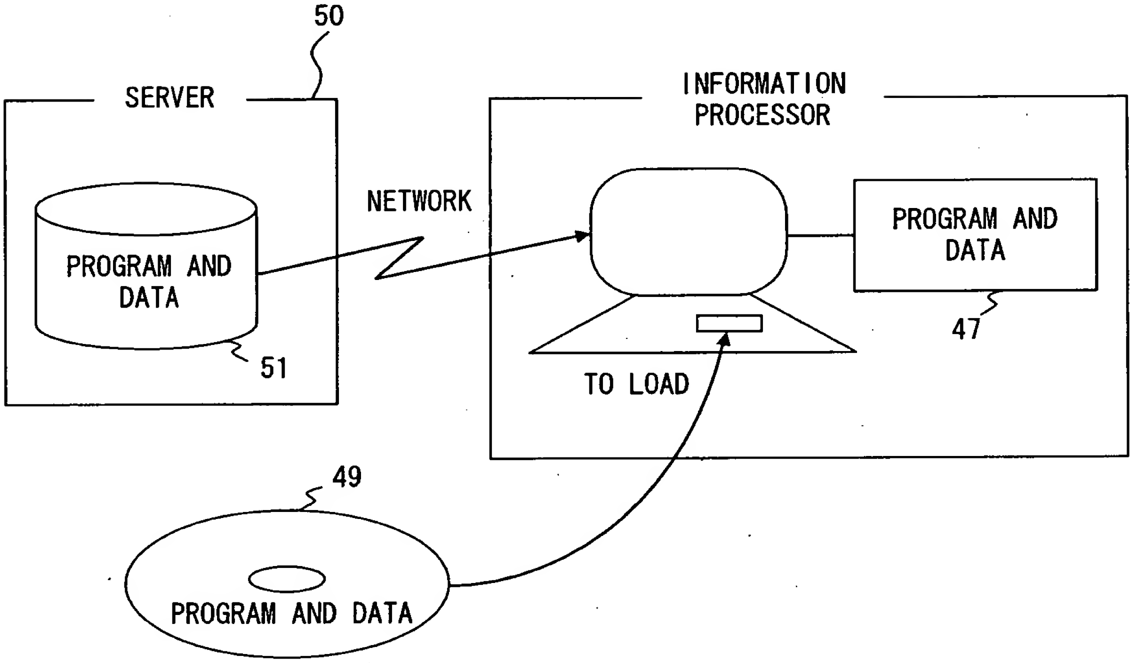
[FIG. 2] BASIC CONFIGURATION OF THE DOCUMENT READING APPARATUS


DOCUMENT TO BE READ
11
12
INPUT UNIT
21
WORD RECOGNITION UNIT
22
MORPHOLOGICAL ANALYSIS UNIT
23
WORD DICTIONARY
24
THEMATIC HIERARCHY RECOGNITION UNIT
25
THEMATIC BOUNDARY CANDIDATE SECTION RECOGNIZING UNIT
26
COMMON TOPIC EXTRACTION UNIT
27
OUTPUT UNIT
28
13
USER

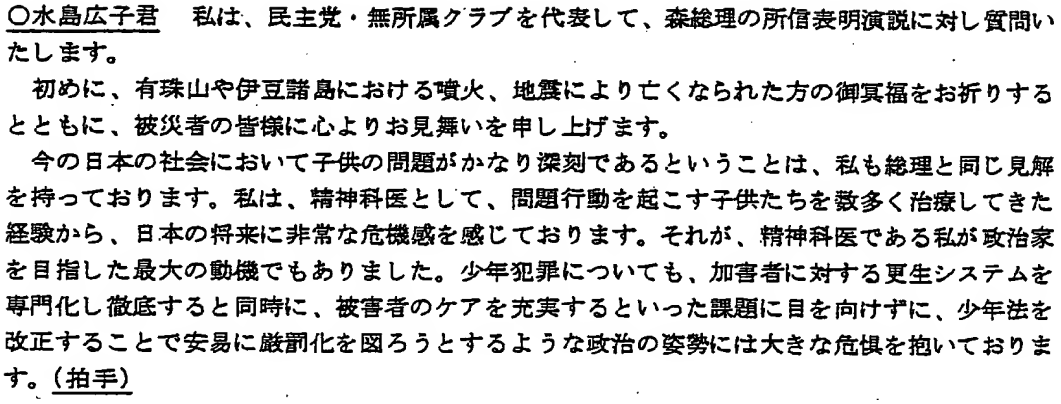
[FIG. 3] CONFIGURATION OF THE INFORMATION PROCESSOR



[FIG. 4] COMPUTER-READABLE STORAGE MEDIUM



[FIG. 5] EXAMPLE OF A DOCUMENT TO BE READ

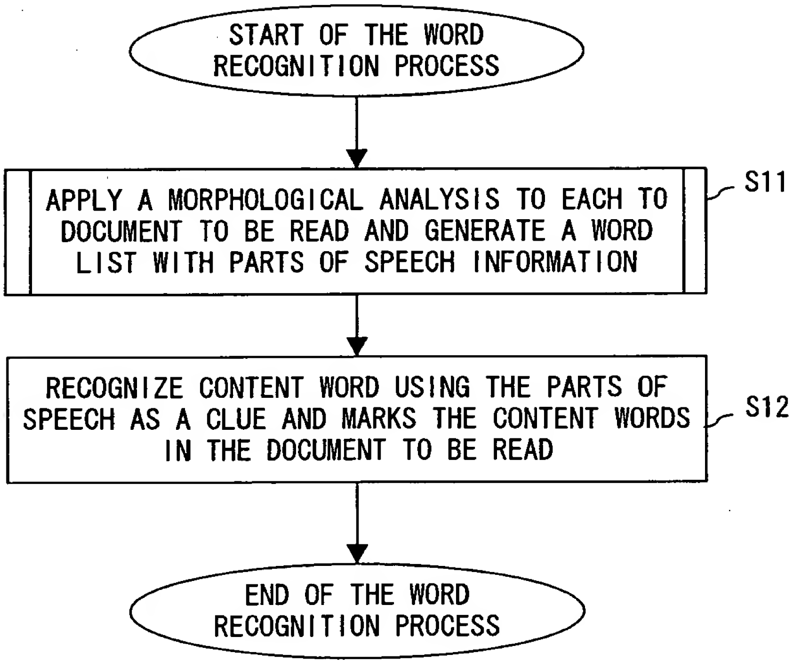


〇水島広子君　私は、民主党・無所属クラブを代表して、森総理の所信表明演説に対し質問いたします。

初めに、有珠山や伊豆諸島における噴火、地震により亡くなられた方の御冥福をお祈りするとともに、被災者の皆様に心よりお見舞いを申し上げます。

今の日本の社会において子供の問題がかなり深刻であるということは、私も総理と同じ見解を持っております。私は、精神科医として、問題行動を起こす子供たちを数多く治療してきた経験から、日本の将来に非常な危機感を感じております。それが、精神科医である私が政治家を目指した最大の動機でもありました。少年犯罪についても、加害者に対する更生システムを専門化し徹底すると同時に、被害者のケアを充実するといった課題に目を向けずに、少年法を改正することで安易に厳罰化を図ろうとするような政治の姿勢には大きな危惧を抱いております。（拍手）

[FIG. 6] DOCUMENT TO BE READ


START OF THE WORD RECOGNITION PROCESS
APPLY A MORPHOLOGICAL ANALYSIS TO EACH TO DOCUMENT TO BE READ AND GENERATE A WORD LIST WITH PARTS OF SPEECH INFORMATION
S11
RECOGNIZE CONTENT WORD USING THE PARTS OF SPEECH AS A CLUE AND MARKS THE CONTENT WORDS IN THE DOCUMENT TO BE READ
S12
END OF THE WORD RECOGNITION PROCESS

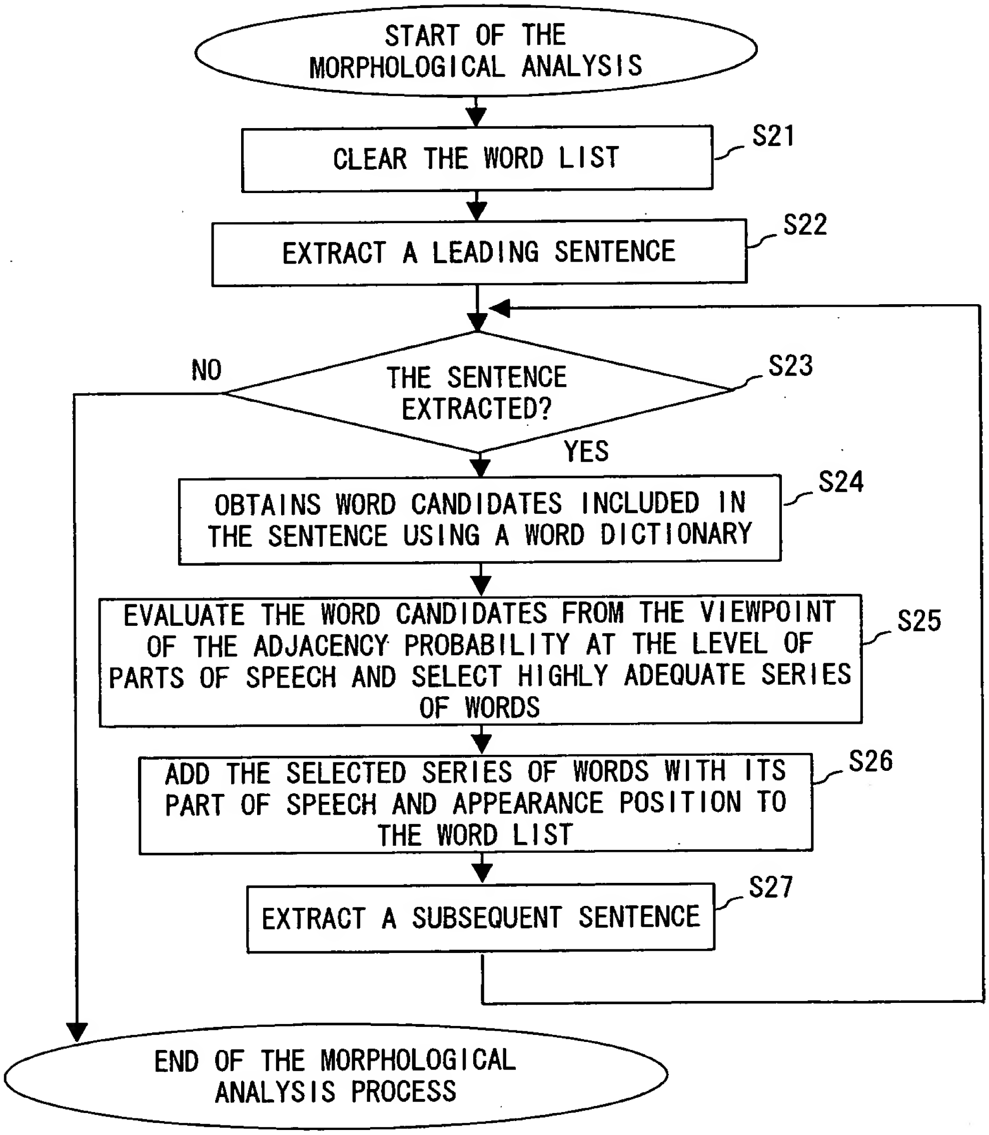
[FIG. 7] EXAMPLE OF THE WORD RECOGNITION RESULT

私は、【民主党/】・【無所属/】【クラブ/】を【代表/する】して、【森/】【総理/】の【所信/】【表明/する】【演説/する】に【対/する】し【質問/する】いたします。

初めに、【有珠山/】や【伊豆/】【諸島/】における【噴火/する】、【地震/】に【よ/る】り【亡くな/る】られた方の【御/する】【冥福/】を【お祈り/する】するとともに、【被災者/】の皆様に【心/】よりお【見舞い/】を【申し上げ/る】ます。

今の【日本/】の【社会/】において【子供/】の【問題/】がかなり【深刻/】であると【い/う】うことは、私も【総理/】と【同じ/】【見解/】を【持/つ】って【お/る】ります。私は、【精神科医/】として、【問題/】【行動/する】を【起こ/す】す【子供/】たちを数多く【治療/する】して【き/】た【経験/する】から、【日本/】の将来に非【常/】な【危機感/】を【感じ/る】て【お/る】ります。それが、【精神科医/】である私が【政治家/】を【目指/す】した【最大/】の【動機/】でもありました。【少年/】【犯罪/】についても、【加害者/】に【対/する】する【更生/する】【システム/】を【専門化/する】し【徹底/する】すると同時に、【被害者/】の【ケア/】を【充実/する】すると【い/う】った【課題/】に【目/】を【向け/る】ずに、【少年/】【法/】を【改正/する】することで【安易/】に【厳罰化/】を【図/】【ろう/】とするような【政治/】の【姿勢/】には大きな【危惧/する】を【抱/く】いて【お/る】ります。

[FIG. 8] FLOWCHART SHOWING THE MORPHOLOGICAL ANALYSIS PROCESS


START OF THE MORPHOLOGICAL ANALYSIS
CLEAR THE WORD LIST
S21
EXTRACT A LEADING SENTENCE
S22
THE SENTENCE EXTRACTED?
S23
NO
YES
OBTAINS WORD CANDIDATES INCLUDED IN THE SENTENCE USING A WORD DICTIONARY
S24
EVALUATE THE WORD CANDIDATES FROM THE VIEWPOINT OF THE ADJACENCY PROBABILITY AT THE LEVEL OF PARTS OF SPEECH AND SELECT HIGHLY ADEQUATE SERIES OF WORDS
S25
ADD THE SELECTED SERIES OF WORDS WITH ITS PART OF SPEECH AND APPEARANCE POSITION TO THE WORD LIST
S26
EXTRACT A SUBSEQUENT SENTENCE
S27
END OF THE MORPHOLOGICAL ANALYSIS PROCESS

[FIG. 9] EXAMPLE HOW TO CONSULT AN ENGLISH DICTIONARY

INPUT SENTENCE 東京都は大都市だ

	TITLE (WORD ROOT)	PART OF SPEECH
CANDIDATE WORD	東	NOUN
	東京都	NOUN
	京都	NOUN
	は	PARTICLE [は]
	大	PREFIX
	都市	NOUN
	だ	AUXILIARY VERB [だ]

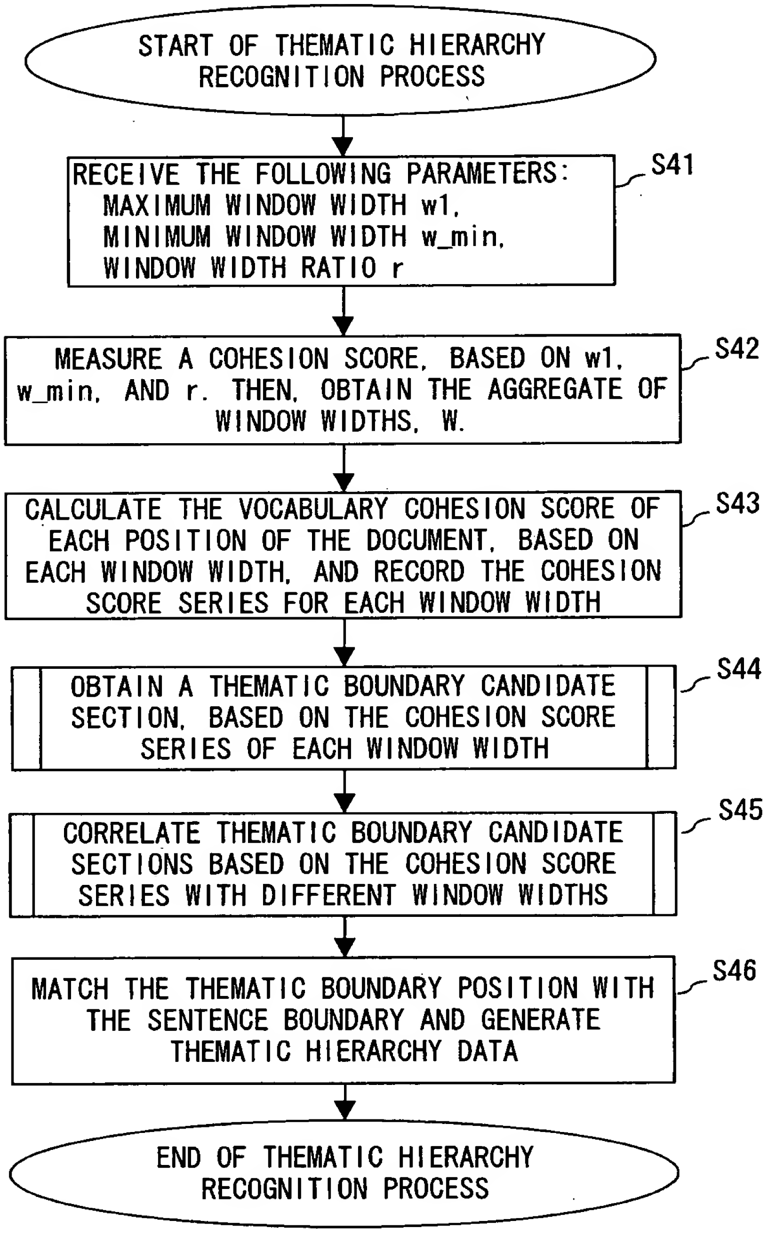
[FIG. 10] EXAMPLE OF HOW TO CONSULT AN ENGLISH DICTIONARY

INPUT SENTENCE Tokyo is the Japanese capital.

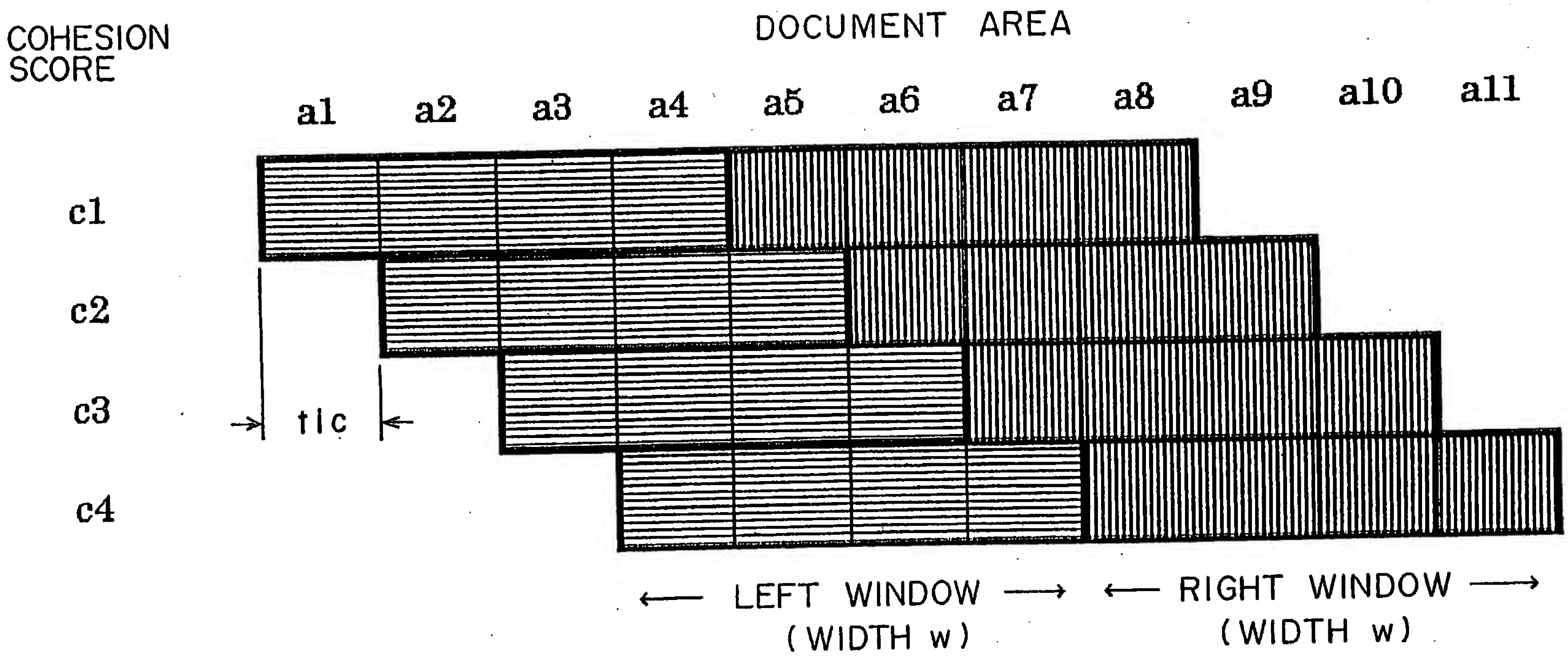
CANDIDATE WORD

headword	base(root) form	part of speech
Tokyo	Tokyo	proper noun
is	be	be verb (the third person singular present form)
the	the	definite article
Japanese	Japanese	proper noun
Japanese	Japanese	adjective
capital	capital	noun

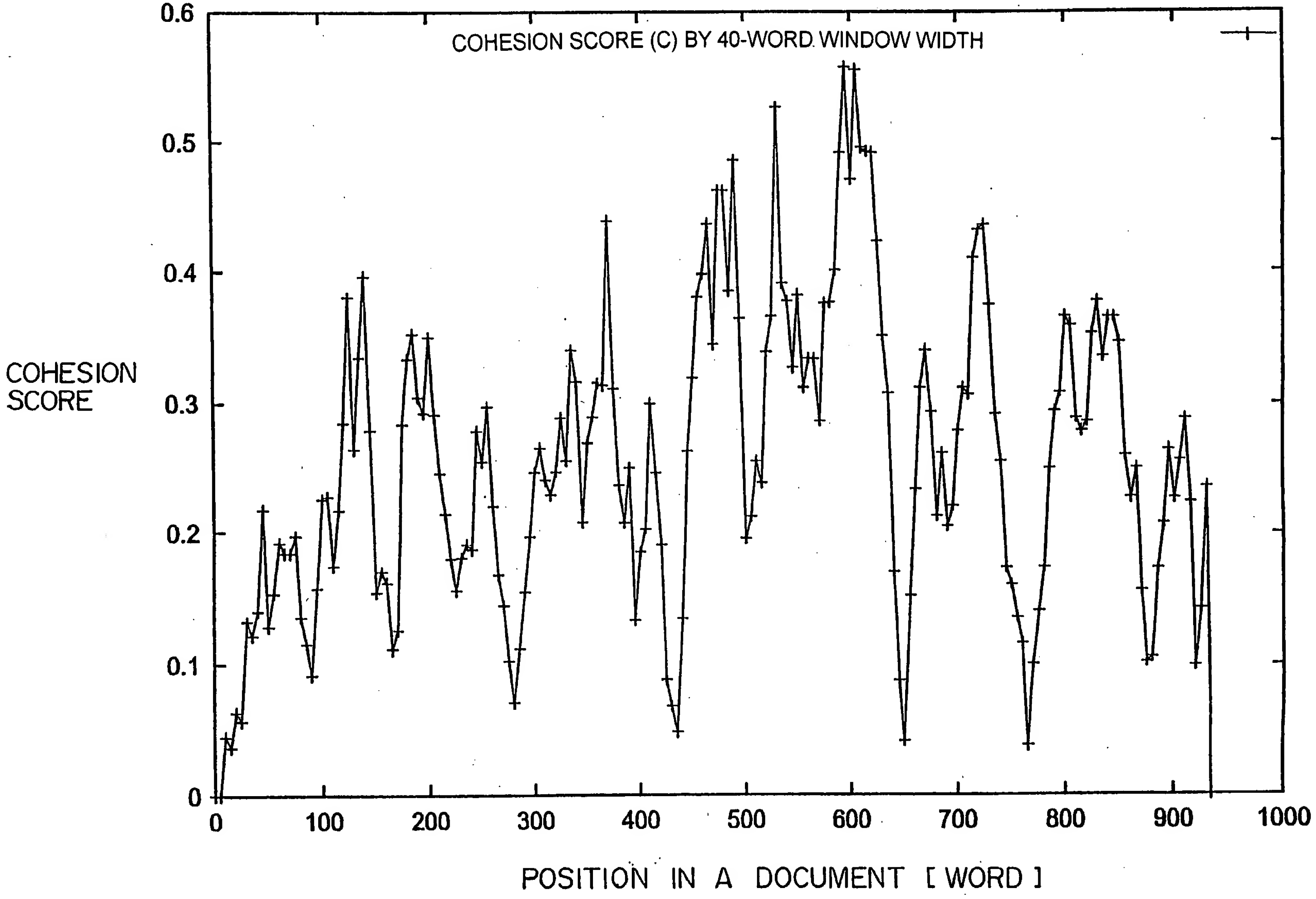
[FIG. 11]FLOWCHART SHOWING THE THEMATIC HIERARCHY RECOGNITION PROCESS


START OF THEMATIC HIERARCHY RECOGNITION PROCESS
RECEIVE THE FOLLOWING PARAMETERS:
MAXIMUM WINDOW WIDTH w1,
MINIMUM WINDOW WIDTH w_min,
WINDOW WIDTH RATIO r
S41
MEASURE A COHESION SCORE, BASED ON w1, w_min, AND r. THEN, OBTAIN THE AGGREGATE OF WINDOW WIDTHS, W.
S42
CALCULATE THE VOCABULARY COHESION SCORE OF EACH POSITION OF THE DOCUMENT, BASED ON EACH WINDOW WIDTH, AND RECORD THE COHESION SCORE SERIES FOR EACH WINDOW WIDTH
S43
OBTAIN A THEMATIC BOUNDARY CANDIDATE SECTION, BASED ON THE COHESION SCORE SERIES OF EACH WINDOW WIDTH
S44
CORRELATE THEMATIC BOUNDARY CANDIDATE SECTIONS BASED ON THE COHESION SCORE SERIES WITH DIFFERENT WINDOW WIDTHS
S45
MATCH THE THEMATIC BOUNDARY POSITION WITH THE SENTENCE BOUNDARY AND GENERATE THEMATIC HIERARCHY DATA
S46
END OF THEMATIC HIERARCHY RECOGNITION PROCESS

[FIG. 12] EXAMPLE OF THE SERIES OF THE COHESION SCORES



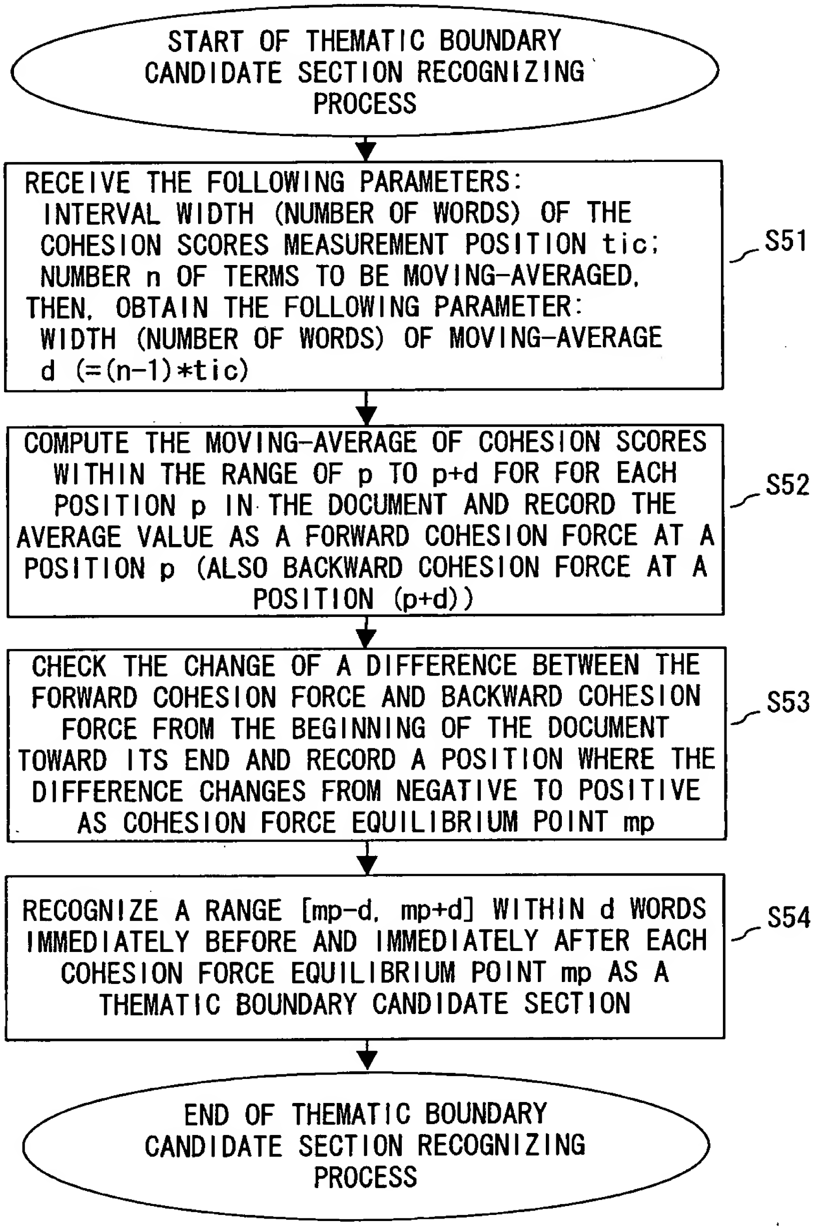
[FIG. 13] EXAMPLE OF THE COHESION SCORE DISTRIBUTION



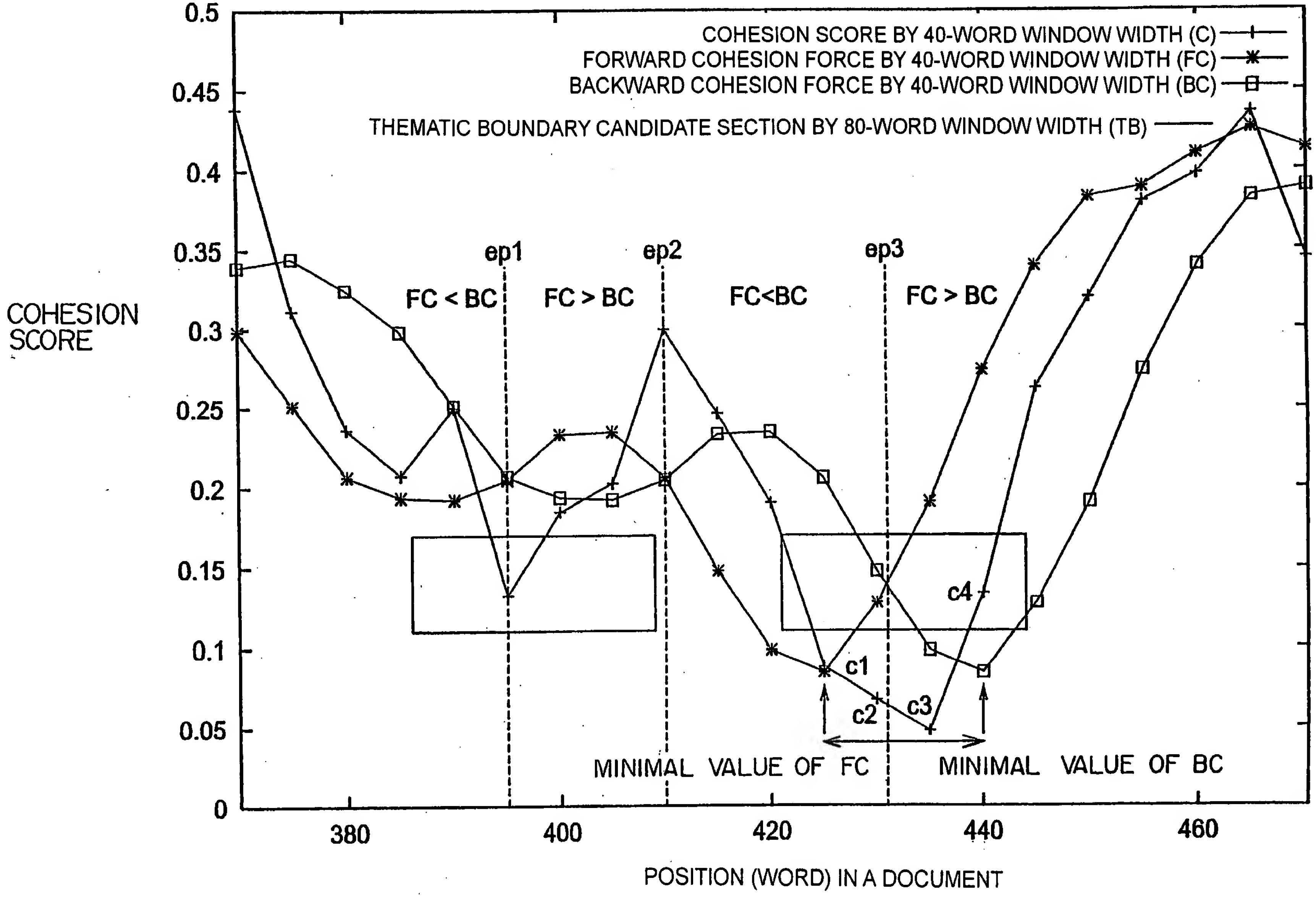
[FIG. 14] CONTRIBUTION OF A DOCUMENT AREA TO THE MOVING AVERAGE

NUMBER OF TERMS		NUMBER OF TIMES TO BE MOVING-AVERAGED										
		a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11
AVERAGE OF FOUR TERMS (c1～c4)	LEFT WINDOW	1	2	3	4	3	2	1	0	0	0	0
	RIGHT WINDOW	0	0	0	0	1	2	3	4	3	2	1
AVERAGE OF THREE TERMS (c1～c3)	LEFT WINDOW	1	2	3	3	2	1	0	0	0	0	
	RIGHT WINDOW	0	0	0	0	1	2	3	3	2	1	
AVERAGE OF TWO TERMS (c1, c2)	LEFT WINDOW	1	2	2	2	1	0	0	0	0		
	RIGHT WINDOW	0	0	0	0	1	2	2	2	1		

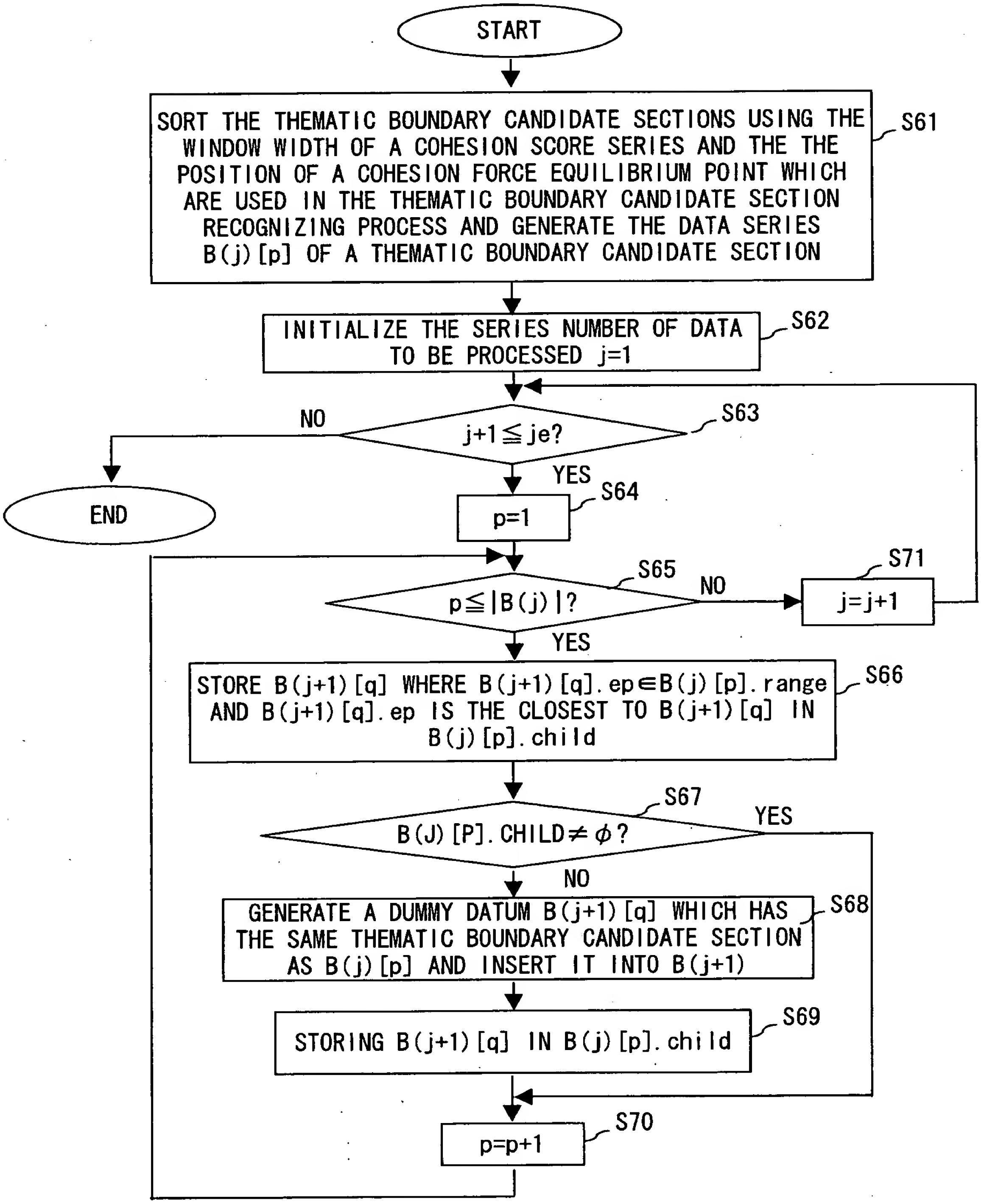
[FIG. 15]
FLOWCHART SHOWING THE THEMATIC BOUNDARY CANDIDATE SECTION RECOGNIZING PROCESS


START OF THEMATIC BOUNDARY CANDIDATE SECTION RECOGNIZING PROCESS
RECEIVE THE FOLLOWING PARAMETERS: INTERVAL WIDTH (NUMBER OF WORDS) OF THE COHESION SCORES MEASUREMENT POSITION tic; NUMBER n OF TERMS TO BE MOVING-AVERAGED. THEN, OBTAIN THE FOLLOWING PARAMETER: WIDTH (NUMBER OF WORDS) OF MOVING-AVERAGE d (=(n-1)*tic)
S51
COMPUTE THE MOVING-AVERAGE OF COHESION SCORES WITHIN THE RANGE OF p TO p+d FOR FOR EACH POSITION p IN THE DOCUMENT AND RECORD THE AVERAGE VALUE AS A FORWARD COHESION FORCE AT A POSITION p (ALSO BACKWARD COHESION FORCE AT A POSITION (p+d))
S52
CHECK THE CHANGE OF A DIFFERENCE BETWEEN THE FORWARD COHESION FORCE AND BACKWARD COHESION FORCE FROM THE BEGINNING OF THE DOCUMENT TOWARD ITS END AND RECORD A POSITION WHERE THE DIFFERENCE CHANGES FROM NEGATIVE TO POSITIVE AS COHESION FORCE EQUILIBRIUM POINT mp
S53
RECOGNIZE A RANGE [mp-d, mp+d] WITHIN d WORDS IMMEDIATELY BEFORE AND IMMEDIATELY AFTER EACH COHESION FORCE EQUILIBRIUM POINT mp AS A THEMATIC BOUNDARY CANDIDATE SECTION
S54
END OF THEMATIC BOUNDARY CANDIDATE SECTION RECOGNIZING PROCESS

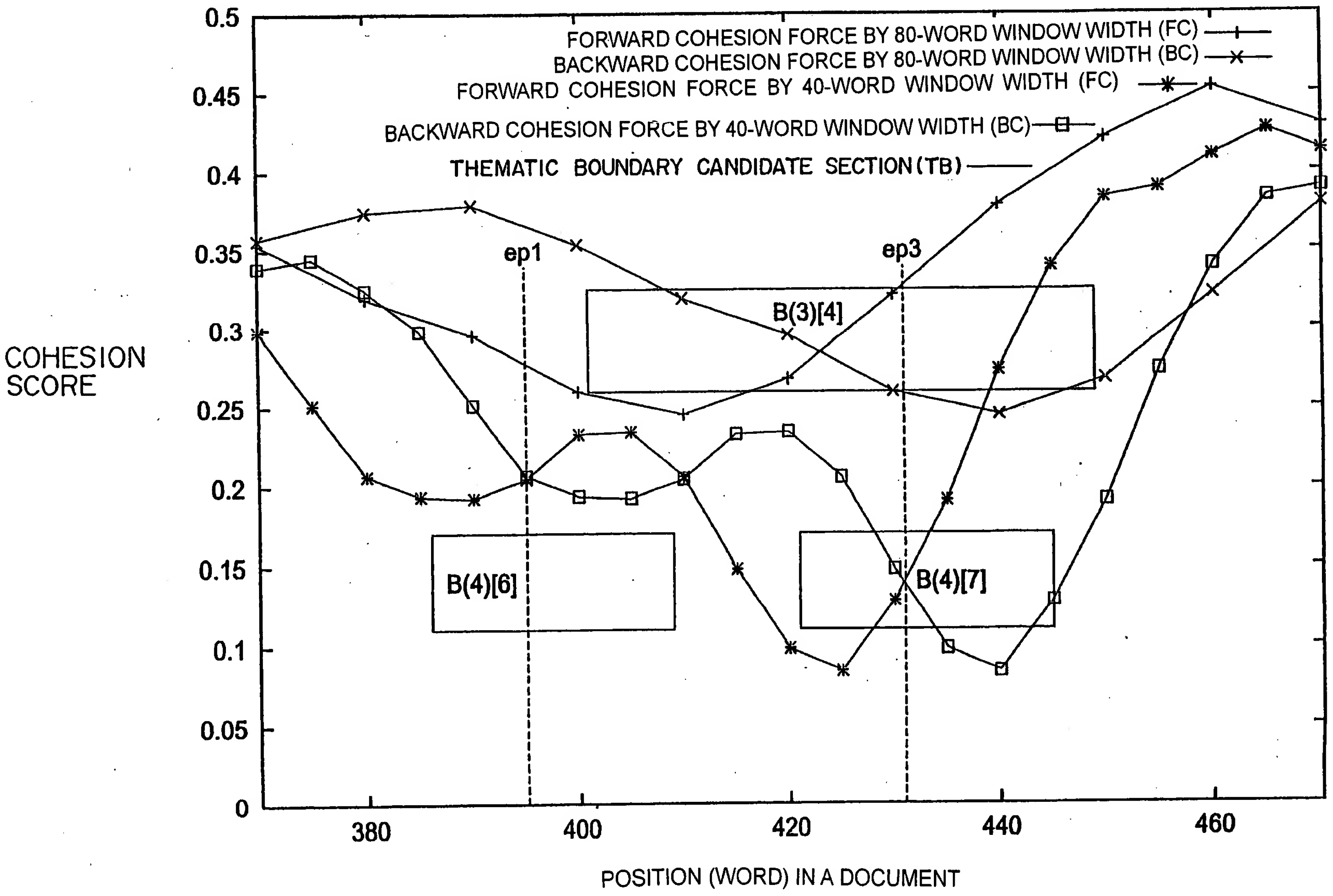
[FIG. 16] EXAMPLE OF THE COHESION FORCE EQUILIBRIUM POINT



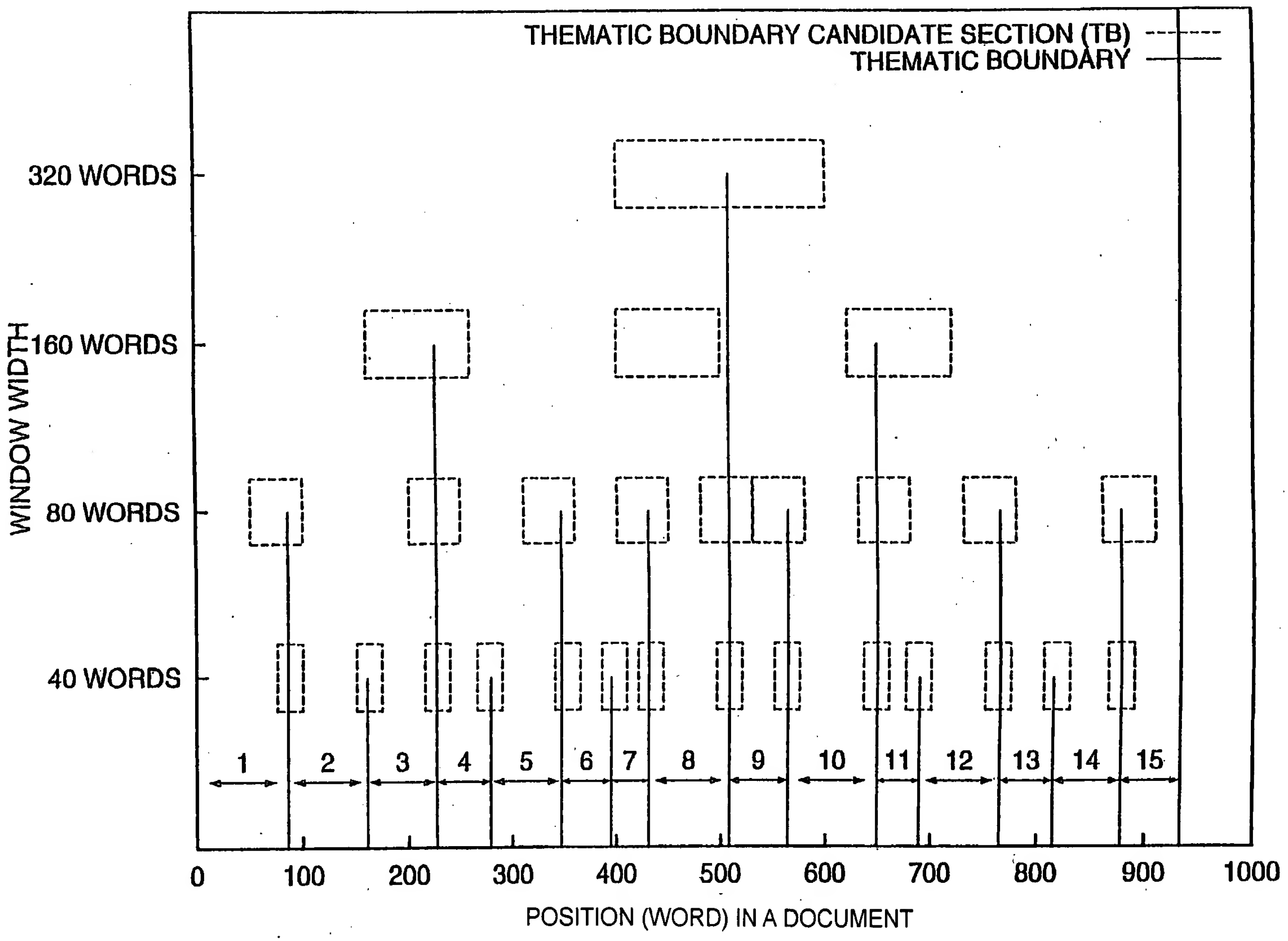
[FIG.17] FLOWCHART SHOWING THE THEMATIC BOUNDARY RECOGNITION PROCESS


START
SORT THE THEMATIC BOUNDARY CANDIDATE SECTIONS USING THE WINDOW WIDTH OF A COHESION SCORE SERIES AND THE THE POSITION OF A COHESION FORCE EQUILIBRIUM POINT WHICH ARE USED IN THE THEMATIC BOUNDARY CANDIDATE SECTION RECOGNIZING PROCESS AND GENERATE THE DATA SERIES B(j)[p] OF A THEMATIC BOUNDARY CANDIDATE SECTION
S61
INITIALIZE THE SERIES NUMBER OF DATA TO BE PROCESSED j=1
S62
j+1≦je?
S63
NO
END
YES
p=1
S64
p≦|B(j)|?
S65
NO
j=j+1
S71
YES
STORE B(j+1)[q] WHERE B(j+1)[q].ep∈B(j)[p].range AND B(j+1)[q].ep IS THE CLOSEST TO B(j+1)[q] IN B(j)[p].child
S66
B(J)[P].CHILD≠φ?
S67
YES
NO
GENERATE A DUMMY DATUM B(j+1)[q] WHICH HAS THE SAME THEMATIC BOUNDARY CANDIDATE SECTION AS B(j)[p] AND INSERT IT INTO B(j+1)
S68
STORING B(j+1)[q] IN B(j)[p].child
S69
p=p+1
S70

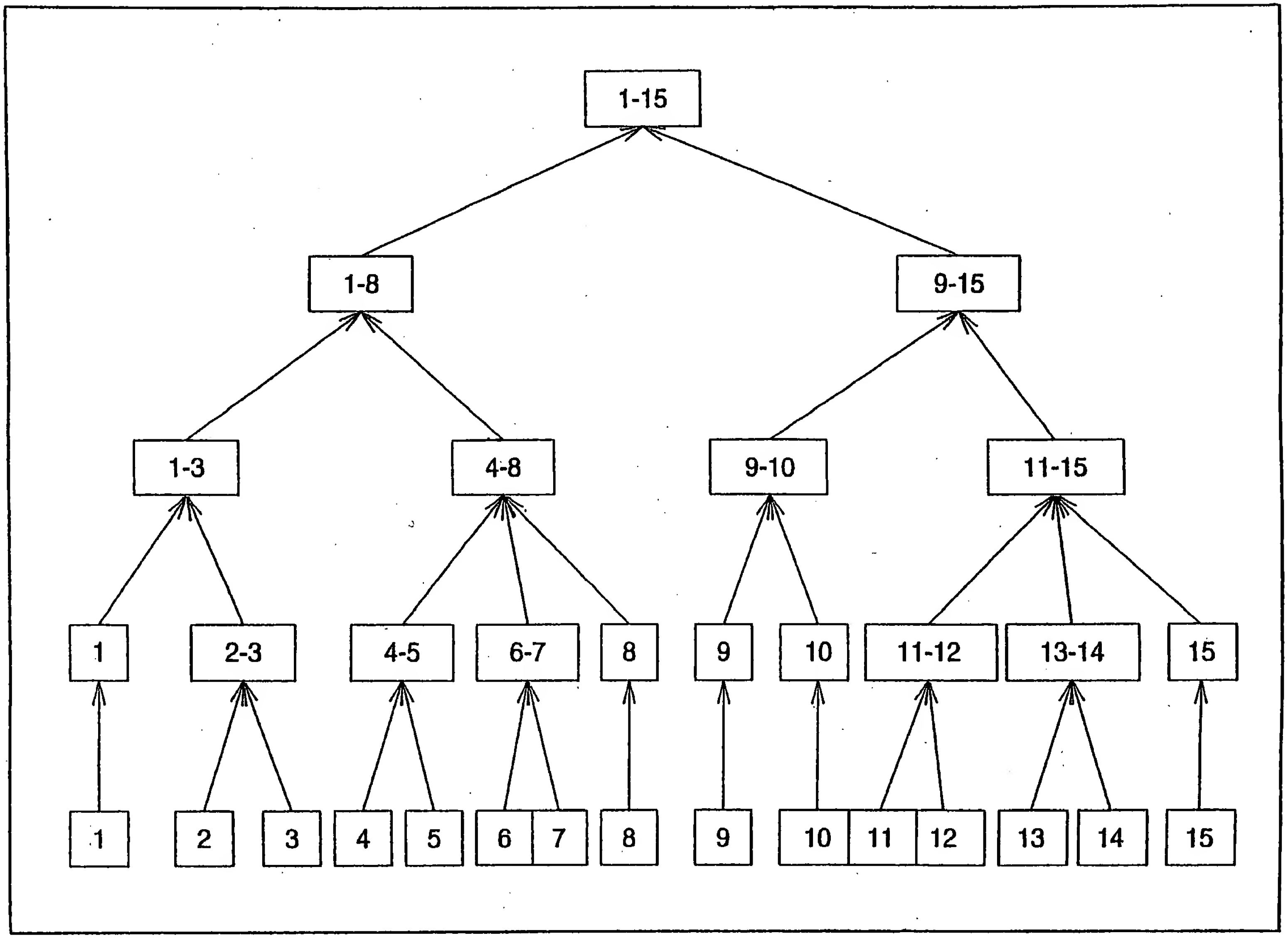
[FIG. 18] EXAMPLE OF DATA TO BE CORRELATED



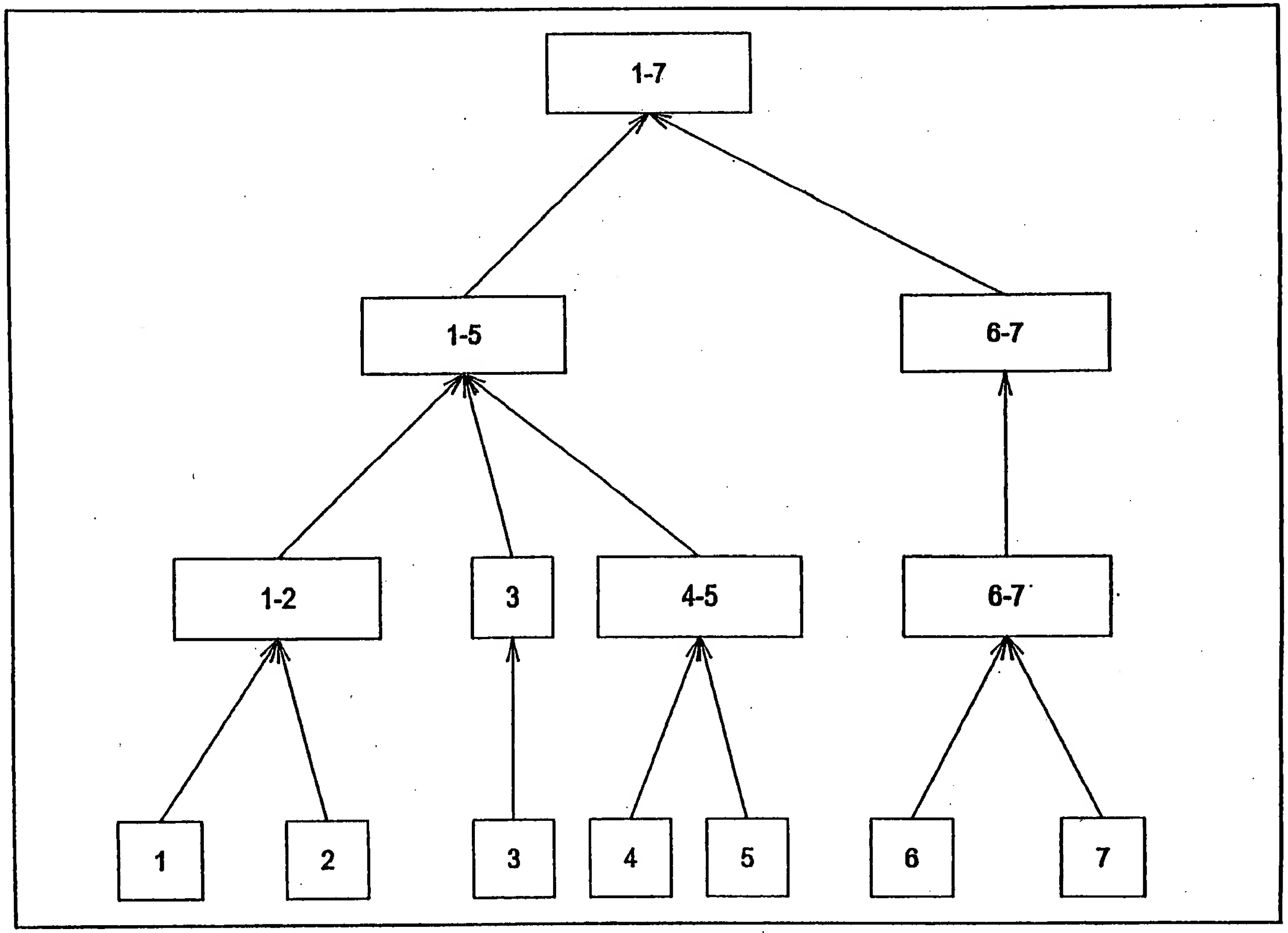
[FIG. 19] EXAMPLE OF THE THEMATIC BOUNDARY RECOGNITION RESULT



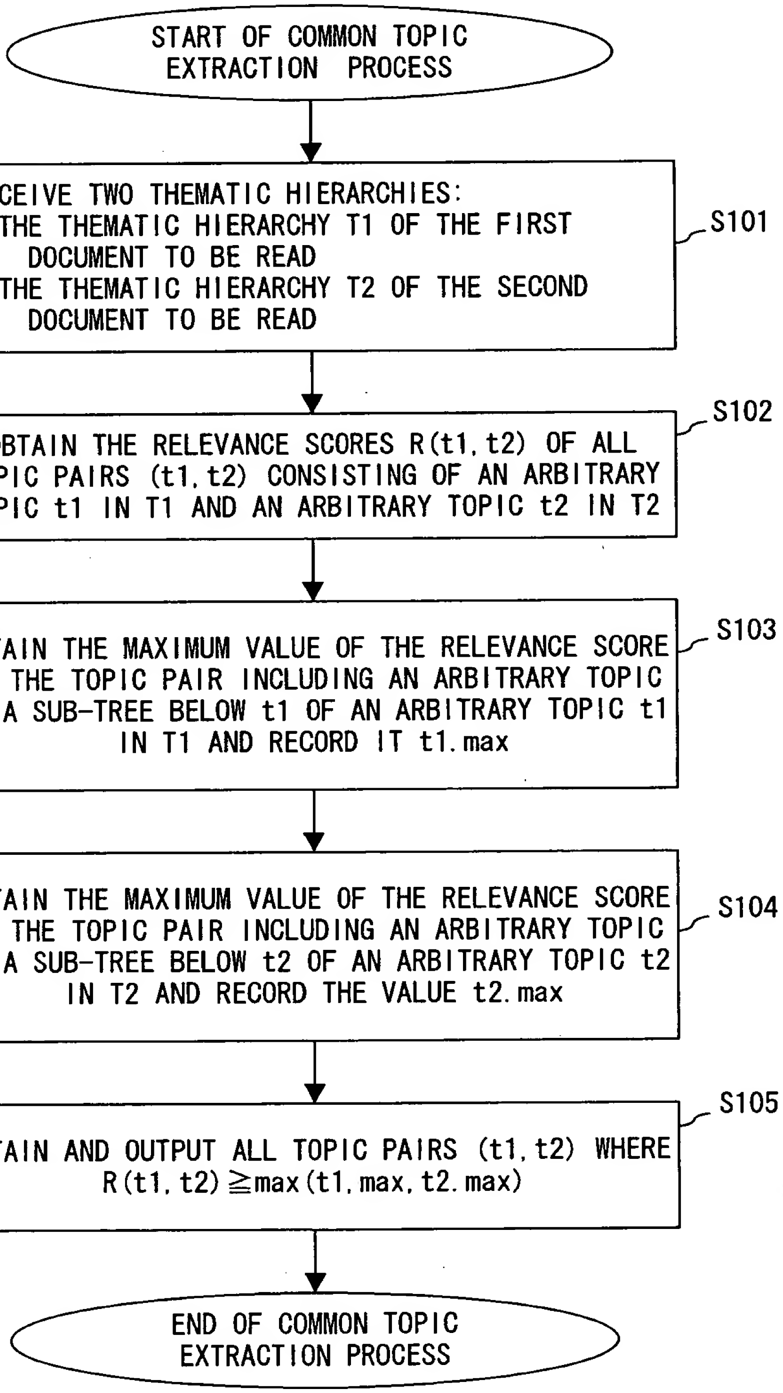
[FIG. 20] EXAMPLE OF THE THEMATIC HIERARCHY OF THE FIRST DOCUMENT TO BE READ



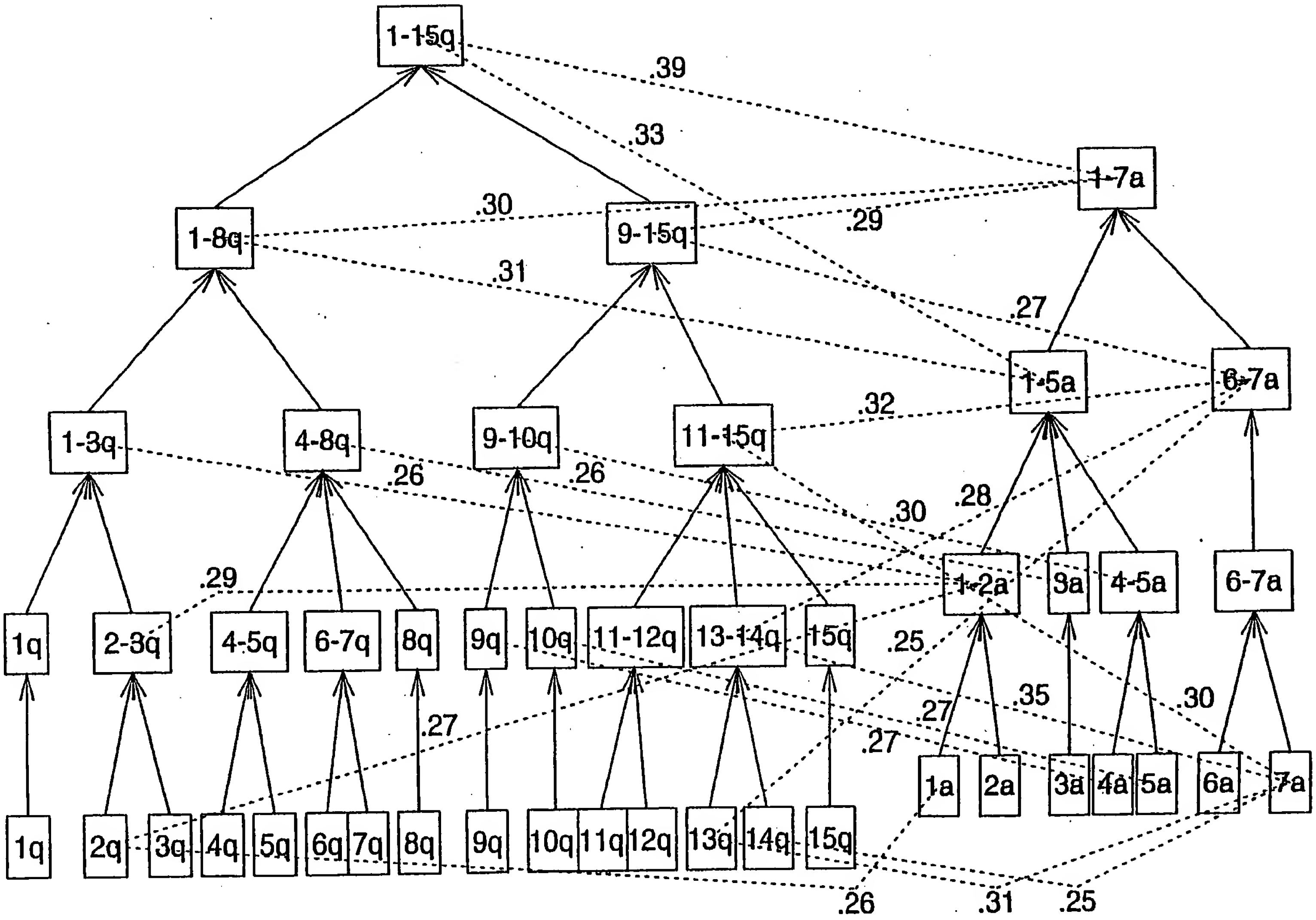
[FIG. 21] EXAMPLE OF THE THEMATIC HIERARCHY OF THE SECOND DOCUMENT TO BE READ



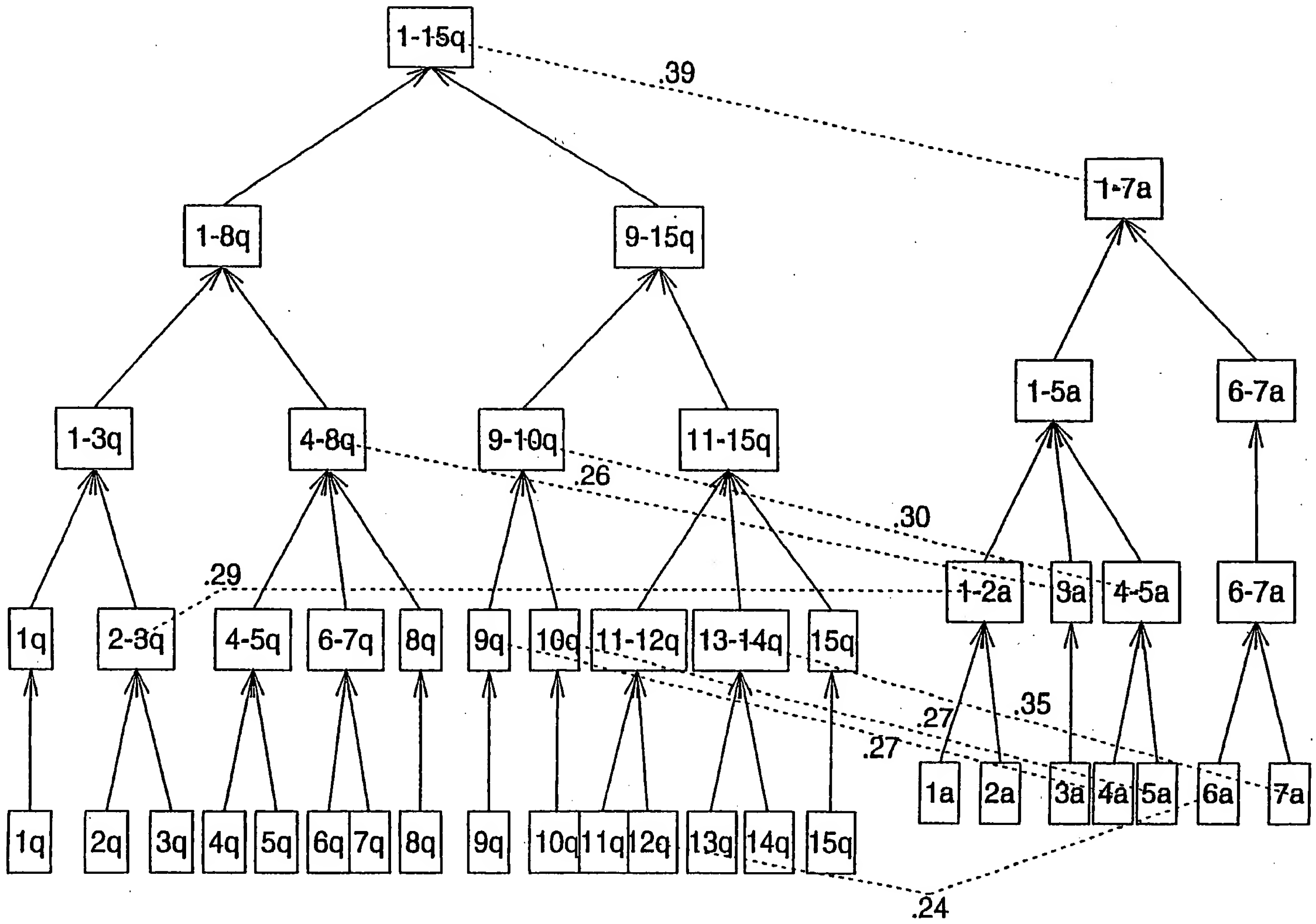
[FIG. 22] FLOWCHART SHOWING THE COMMON TOPIC EXTRACTION PROCESS


START OF COMMON TOPIC EXTRACTION PROCESS
RECEIVE TWO THEMATIC HIERARCHIES:
THE THEMATIC HIERARCHY T1 OF THE FIRST DOCUMENT TO BE READ
THE THEMATIC HIERARCHY T2 OF THE SECOND DOCUMENT TO BE READ
S101
OBTAIN THE RELEVANCE SCORES R(t1, t2) OF ALL TOPIC PAIRS (t1, t2) CONSISTING OF AN ARBITRARY TOPIC t1 IN T1 AND AN ARBITRARY TOPIC t2 IN T2
S102
OBTAIN THE MAXIMUM VALUE OF THE RELEVANCE SCORE IN THE TOPIC PAIR INCLUDING AN ARBITRARY TOPIC IN A SUB-TREE BELOW t1 OF AN ARBITRARY TOPIC t1 IN T1 AND RECORD IT t1. max
S103
OBTAIN THE MAXIMUM VALUE OF THE RELEVANCE SCORE IN THE TOPIC PAIR INCLUDING AN ARBITRARY TOPIC IN A SUB-TREE BELOW t2 OF AN ARBITRARY TOPIC t2 IN T2 AND RECORD THE VALUE t2. max
S104
OBTAIN AND OUTPUT ALL TOPIC PAIRS (t1, t2) WHERE R(t1, t2) ≧max(t1. max, t2. max)
S105
END OF COMMON TOPIC EXTRACTION PROCESS

[FIG. 23] EXAMPLE OF RELEVANCE SCORE CALCULATION RESULT



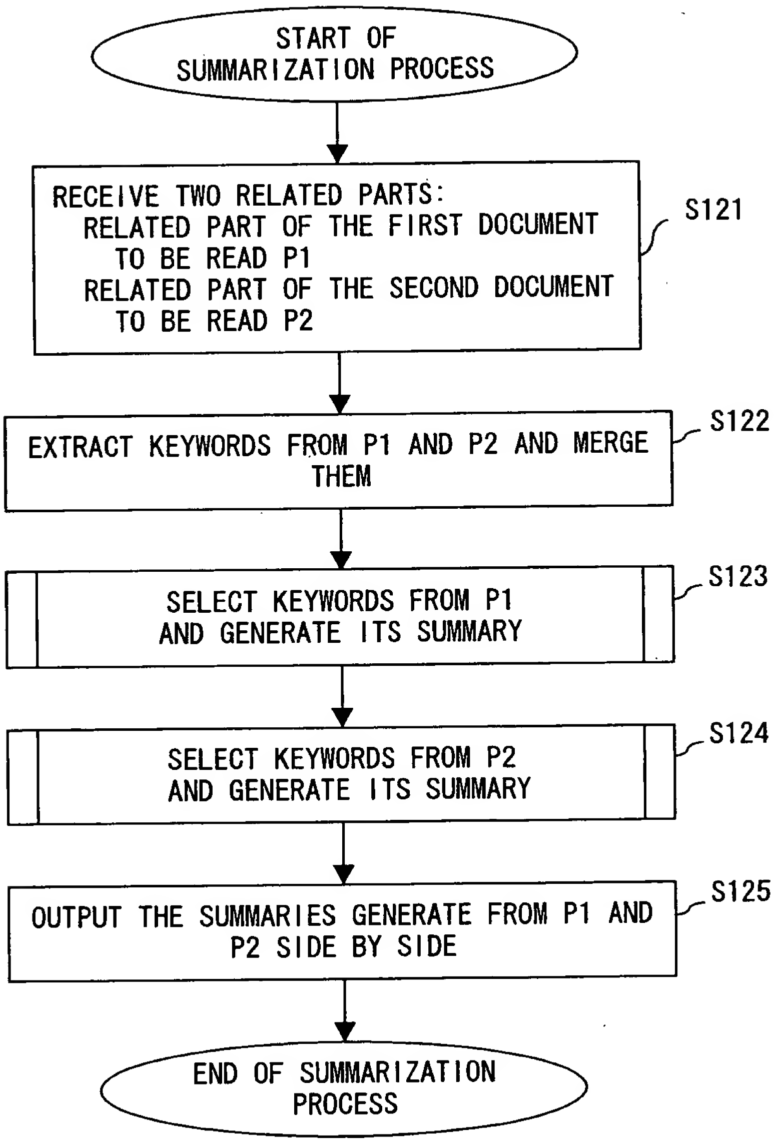
[FIG. 24] EXAMPLE OF COMMON TOPIC EXTRACTION RESULT



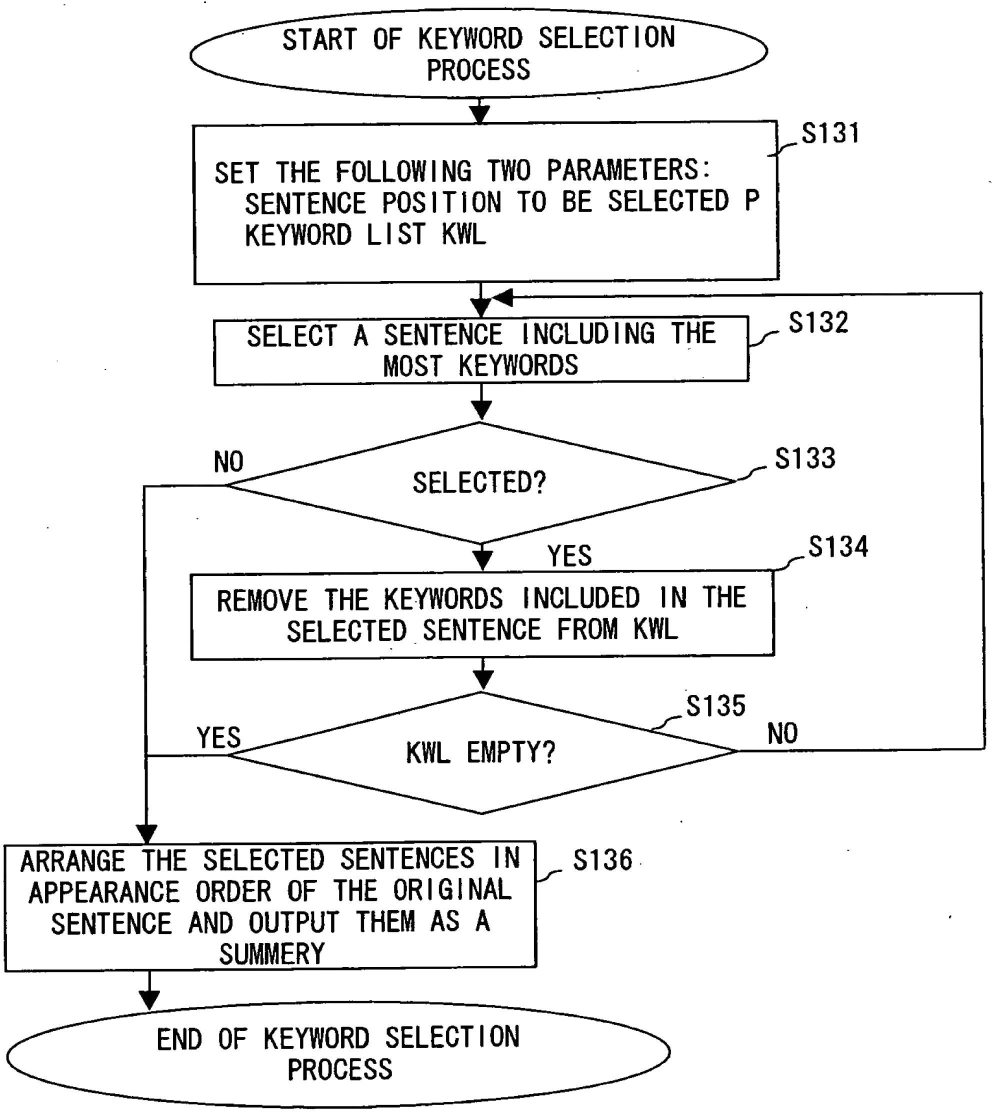
[FIG. 25] EXAMPLE OF THE OUTPUT RELATED PARTS

○水島広子君 [9-10]	○内閣総理大臣（森喜朗君）[4-5]
¶9 総理御自身も触れられている大人社会のあり方ですが、これが子供たちに大きな影響を与えるのは事実だと思います。子供たちは大人のまねをして成長します。大人社会のモラルがこれほど低下した今の日本で、子供たちのモラルだけが高まったら、むしろおかしなことだと思います。モラルの低下の一つの例として、子供の目に触れるテレビや雑誌、ゲームなどの影響も無視できません。だれでも簡単に目にするメディアに暴力や性暴力がはんらんし、町じゅうに売春情報があふれているというのが今の大人の社会です。子供たちを批判する前に、総理御自身も含めて、私たち大人がまず反省すべきではないでしょうか。	¶4 テレビや雑誌、ゲームなどの青少年を取り巻く環境について、暴力や性犯罪がはんらんしており、青少年にとって大きな問題であるとの御指摘でありますが、これらの問題は、申すまでもなく大人社会の責任であります。青少年を取り巻く社会環境の改善のため、社会が一体となった取り組みを進めることが極めて重要であると考えております。
¶10 子供たちの問題行動とメディアによる有害情報の関係を指摘する専門家はたくさんいます。仮に犯罪に直結しなくても、幼いころから有害情報に当たり前のように触れることが子供たちの精神面の発育に及ぼす影響は無視できません。諸外国でも進められているように、子供たちを有害な情報から守る法律を日本でも早急につくる必要があると思います。これはもちろん、国家による検閲というような形をとるべきではありません。例えば、子供にとって有害な情報であるか否かを親が判断して選べるようなシステム、また、町中でも子供が有害情報に触れるのを防ぐような社会的なバリアをつくるなど、地域社会の大人たちが子供たちを守るようなシステムをつくるべきだと思います。子供を有害情報から守るための立法の必要性について、森総理はいかがお考えでしょうか。	¶5 また、子供たちを有害情報から守るための法律の早急な制定を促す御意見をいただきました。私は、かねてから、少年非行対策は与野党対立案件にあらずと考えておりますが、御指摘の点については、まさに議員と意見を一にするものであります。しかしながら、この種法律の制定につきましては、青少年をめぐる環境の浄化の基本的なあり方や表現の自由とのかかわりなど、国民的な合意の形成が必要であると考えられ、関係方面の幅広い議論を重ねていきたいと考えております。

[FIG. 26] FLOWCHART SHOWING THE SUMMARIZATION PROCESS


START OF SUMMARIZATION PROCESS
RECEIVE TWO RELATED PARTS: RELATED PART OF THE FIRST DOCUMENT TO BE READ P1 RELATED PART OF THE SECOND DOCUMENT TO BE READ P2
S121
EXTRACT KEYWORDS FROM P1 AND P2 AND MERGE THEM
S122
SELECT KEYWORDS FROM P1 AND GENERATE ITS SUMMARY
S123
SELECT KEYWORDS FROM P2 AND GENERATE ITS SUMMARY
S124
OUTPUT THE SUMMARIES GENERATE FROM P1 AND P2 SIDE BY SIDE
S125
END OF SUMMARIZATION PROCESS

[FIG. 27] FLOWCHART SHOWING THE IMPORTANCE SENTENCE SELECTION PROCESS


START OF KEYWORD SELECTION PROCESS
SET THE FOLLOWING TWO PARAMETERS:
SENTENCE POSITION TO BE SELECTED P
KEYWORD LIST KWL
S131
SELECT A SENTENCE INCLUDING THE MOST KEYWORDS
S132
SELECTED?
S133
NO
YES
REMOVE THE KEYWORDS INCLUDED IN THE SELECTED SENTENCE FROM KWL
S134
KWL EMPTY?
S135
YES
NO
ARRANGE THE SELECTED SENTENCES IN APPEARANCE ORDER OF THE ORIGINAL SENTENCE AND OUTPUT THEM AS A SUMMERY
S136
END OF KEYWORD SELECTION PROCESS

[FIG. 28] EXAMPLE OF THE SUMMARIZED RELATED PARTS (NO. 1)

○水島広子君 [2-3] …精神科医として現場で子供たちや親たちと向き合ってきた私の目には、総理のおっしゃるような教育改革で問題が解決できるとはとても思えません。…いじめの問題を根本的に解決するには、人間の多様性を尊重して、自分も他人も大切にできる子供を育てる教育が不可欠です。…教育基本法の見直しにしても、本来、他者との触れ合いを通して自発的に育てるはずの奉仕の精神や道徳心といったものを法改正によって一方的に押しつけようとするのであれば、逆効果となり、取り返しがつかないことになると思います。…	○内閣総理大臣（森喜朗君）[1-2] …いじめについてのお尋ねでありますが、いじめ問題を解決するためには、弱い者をいじめることは人間として絶対に許されないとの認識のもと、奉仕活動や自然体験活動などを通じて、子供たちに、命を大切にし、他人を思いやる心など、基本的な倫理観をはぐくむことが重要であると考えております。…教育基本法についてのお尋ねでありますが、教育基本法は、制定以来半世紀を経ております。…今後、教育改革国民会議において、例えば我が国の文化や伝統を尊重する気持ちを養う観点や生涯学習時代を迎える観点、あるいは教育において家庭や地域が果たすべき役割といった観点を初め、さまざまな観点から幅広く議論を進めていただきたいと考えております。民法改正についてのお尋ねがありました。

○水島広子君 [4-8] …夫婦は皆同じ名字にするというのは、明治維新に西洋のまねをして導入された制度であり、日本独自の伝統とは何ら関係がありません。…男女共同参画社会の実現を所信表明演説でもうたわれた総理は、希望する夫婦には夫婦別姓を認めるよう民法を直ちに改正することについて、どのようにお考えでしょうか、お尋ねいたします。…非嫡出子は、法律上相続のときに差別を受けるだけではなく、普通に社会生活を送る上でも、就職や結婚の際に差別を受けています。…	○内閣総理大臣（森喜朗君）[3] …この導入についてのお話でありますが、これは婚姻制度や家族のあり方とも関連する重要な問題でありまして、国民や関係各方面の意見が現在分かれている状況にありますので、国民各層の御意見を幅広く聞き、また、各方面における議論の推移も踏まえながら、適切に対処していく必要があるのではないかと考えております。また、嫡出でない子の法定相続分等についてでありますが、民法上、嫡出でない子と嫡出である子の相続分に差異が設けられている点の解消につきましても、選択的夫婦別氏の問題と同様に、国民や関係各方面の意見が分かれている状況にありますので、国民各界各層の御意見をこれも幅広く聞くなどいたしまして、適切に対処していく必要があると考えております。

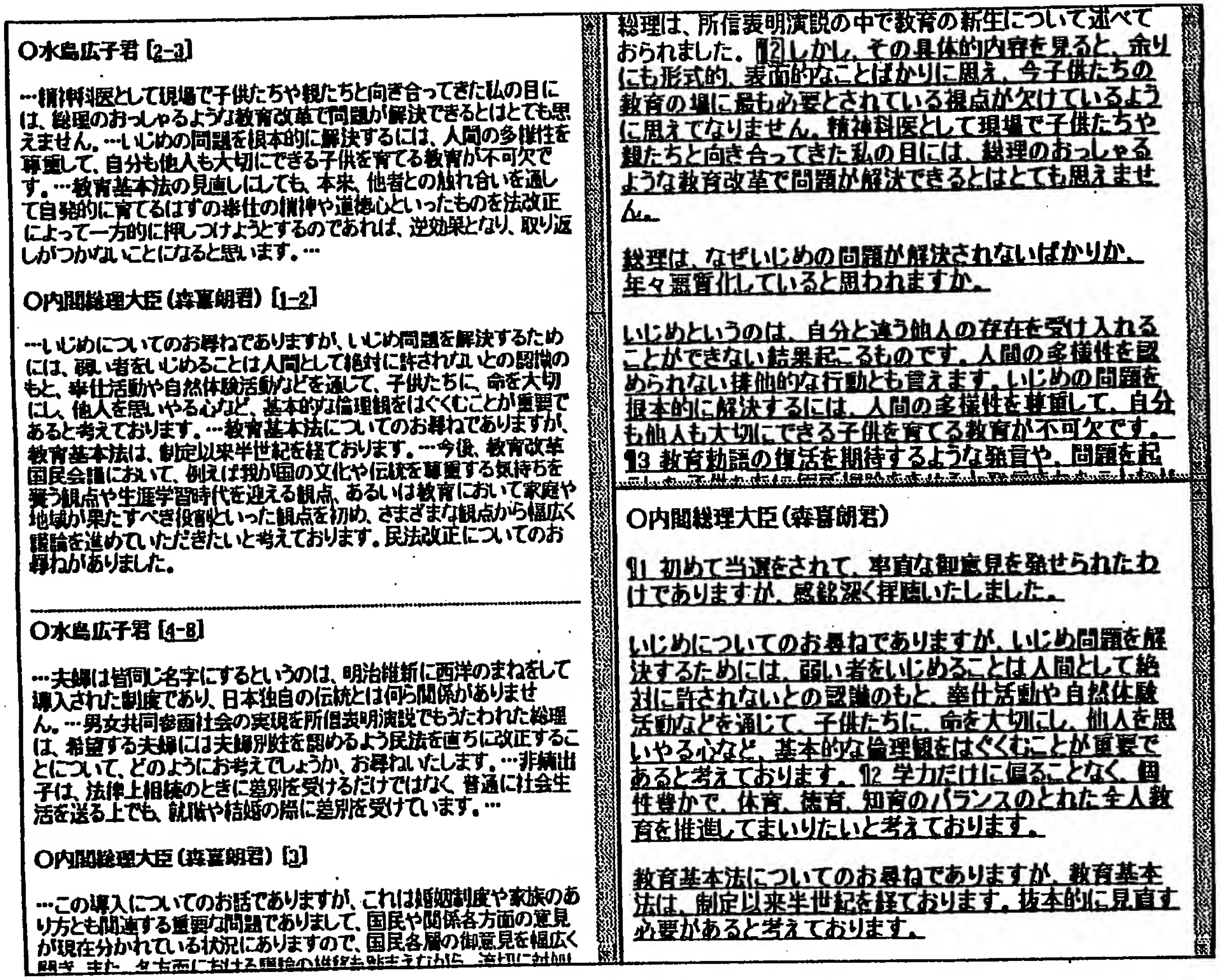
[FIG. 29] EXAMPLE OF THE SUMMARIZED RELATED PARTS (NO. 2)

○水島広子君 [9-10]	○内閣総理大臣（森喜朗君）[4-5]
¶9 総理御自身も触れられている大人社会のあり方ですが、これが子供たちに大きな影響を与えるのは事実だと思います。…モラルの低下の一つの例として、子供の目に触れるテレビや雑誌、ゲームなどの影響も無視できません。だれでも簡単に目にするメディアに暴力や性暴力がはんらんし、町じゅうに売春情報があふれているというのが今の大人の社会です。…	¶4 テレビや雑誌、ゲームなどの青少年を取り巻く環境について、暴力や性犯罪がはんらんしており、青少年にとって大きな問題であるとの御指摘でありますが、これらの問題は、申すまでもなく大人社会の責任であります。…
¶10 …諸外国でも進められているように、子供たちを有害な情報から守る法律を日本でも早急につくる必要があると思います。…	¶5 また、子供たちを有害情報から守るための法律の早急な制定を促す御意見をいただきました。…

[FIG. 30] EXAMPLE OF THE SUMMARIZED RELATED PARTS (NO. 3)

○水島広子君 [12] …しかし、最近の国会運営を見て、民主主義とは話し合いよりも多数決で押し切ることだという誤った理解をしている人たちがふえているように思います。…まずは党首討論を毎週行うことで総理みずから模範を示されるべきだと思いますし、あわせて、各大臣についても同様な場を設けるべきだと思いますが、総理のお考えをお願いいたします。	○内閣総理大臣（森喜朗君）[6] …積極的に議論を闘わせることは、最終的には多数決で決するにせよ、国民の前に争点を明らかにし、国民の政治への関心を高めるために重要なことであると考えます。お尋ねの党首討論のあり方などの国会の運営に関する問題につきましては、国会において御議論をいただきたいと考えます。
○水島広子君 [13-14] …小児科救急の不備のために命を落とした不幸な子供の例も多数報道されています。また、心を病む子供に対しては、大人を中心とした医療体系では対応し切れません。小児の医療は、国の将来を担う貴重な人材を社会全体で守るという発想で行わなければいけないと思います。国が特別な予算枠を確保し良質な医療を提供すること、また、数少ない小児科医を効率的に活用するためにも、各都道府県に子供病院をつくり、人材を集中させ、包括的、専門的な子供医療ができるように国としても取り組むべきだと思いますが、いかがでしょうか。現行のように各都道府県の自助努力に任せていると、いまだに全国で十九都道府県にしか子供の総合医療施設がありません。…	○内閣総理大臣（森喜朗君）[7] …このため、小児専門の救急医療体制の整備、診療報酬における小児医療の適切な評価、小児医療施設の整備の補助などを行っているところでありますが、今後とも適切な対応をしていきたいと考えております。…

[FIG. 31] EXAMPLE OF THE PRESENTED RELATED PARTS WITH AN ORIGINAL DOCUMENT REFERENCE FUNCTION



○水島広子君 [2-3]

…精神科医として現場で子供たちや親たちと向き合ってきた私の目には、総理のおっしゃるような教育改革で問題が解決できるとはとても思えません。…いじめの問題を根本的に解決するには、人間の多様性を尊重して、自分も他人も大切にできる子供を育てる教育が不可欠です。…教育基本法の見直しにしても、本来、他者との触れ合いを通して自発的に育てるはずの奉仕の精神や道徳心といったものを法改正によって一方的に押しつけようとするのであれば、逆効果となり、取り返しがつかないことになると思います。…

○内閣総理大臣（森喜朗君）[1-2]

…いじめについてのお尋ねでありますが、いじめ問題を解決するためには、弱い者をいじめることは人間として絶対に許されないとの認識のもと、奉仕活動や自然体験活動などを通して、子供たちに、命を大切にし、他人を思いやる心など、基本的な倫理観をはぐくむことが重要であると考えております。…教育基本法についてのお尋ねでありますが、教育基本法は、制定以来半世紀を経ております。…今後、教育改革国民会議において、例えば我が国の文化や伝統を尊重する気持ちを養う観点や生涯学習時代を迎える観点、あるいは教育において家庭や地域が果たすべき役割といった観点を初め、さまざまな観点から幅広く議論を進めていただきたいと考えております。民法改正についてのお尋ねがありました。

○水島広子君 [4-8]

…夫婦は皆同じ名字にするというのは、明治維新に西洋のまねをして導入された制度であり、日本独自の伝統とは何ら関係がありません。…男女共同参画社会の実現を所信表明演説でもうたわれた総理は、希望する夫婦には夫婦別姓を認めるよう民法を直ちに改正することについて、どのようにお考えでしょうか、お尋ねいたします。…非嫡出子は、法律上相続のときに差別を受けるだけではなく、普通に社会生活を送る上でも、就職や結婚の際に差別を受けています。…

○内閣総理大臣（森喜朗君）[3]

…この導入についてのお話でありますが、これは婚姻制度や家族のあり方とも関連する重要な問題でありまして、国民や関係各方面の意見が現在分かれている状況にありますので、国民各層の御意見を幅広く聞き、また、各方面における議論の推移を踏まえながら、適切に対処

総理は、所信表明演説の中で教育の新生について述べておられました。¶2 しかし、その具体的内容を見ると、余りにも形式的、表面的なことばかりに思え、今子供たちの教育の場に最も必要とされている視点が欠けているように思えてなりません。精神科医として現場で子供たちや親たちと向き合ってきた私の目には、総理のおっしゃるような教育改革で問題が解決できるとはとても思えません。

総理は、なぜいじめの問題が解決されないばかりか、年々悪質化していると思われますか。

いじめというのは、自分と違う他人の存在を受け入れることができない結果起こるものです。人間の多様性を認められない排他的な行動とも言えます。いじめの問題を根本的に解決するには、人間の多様性を尊重して、自分も他人も大切にできる子供を育てる教育が不可欠です。¶3 教育勅語の復活を期待するような発言や、問題を起

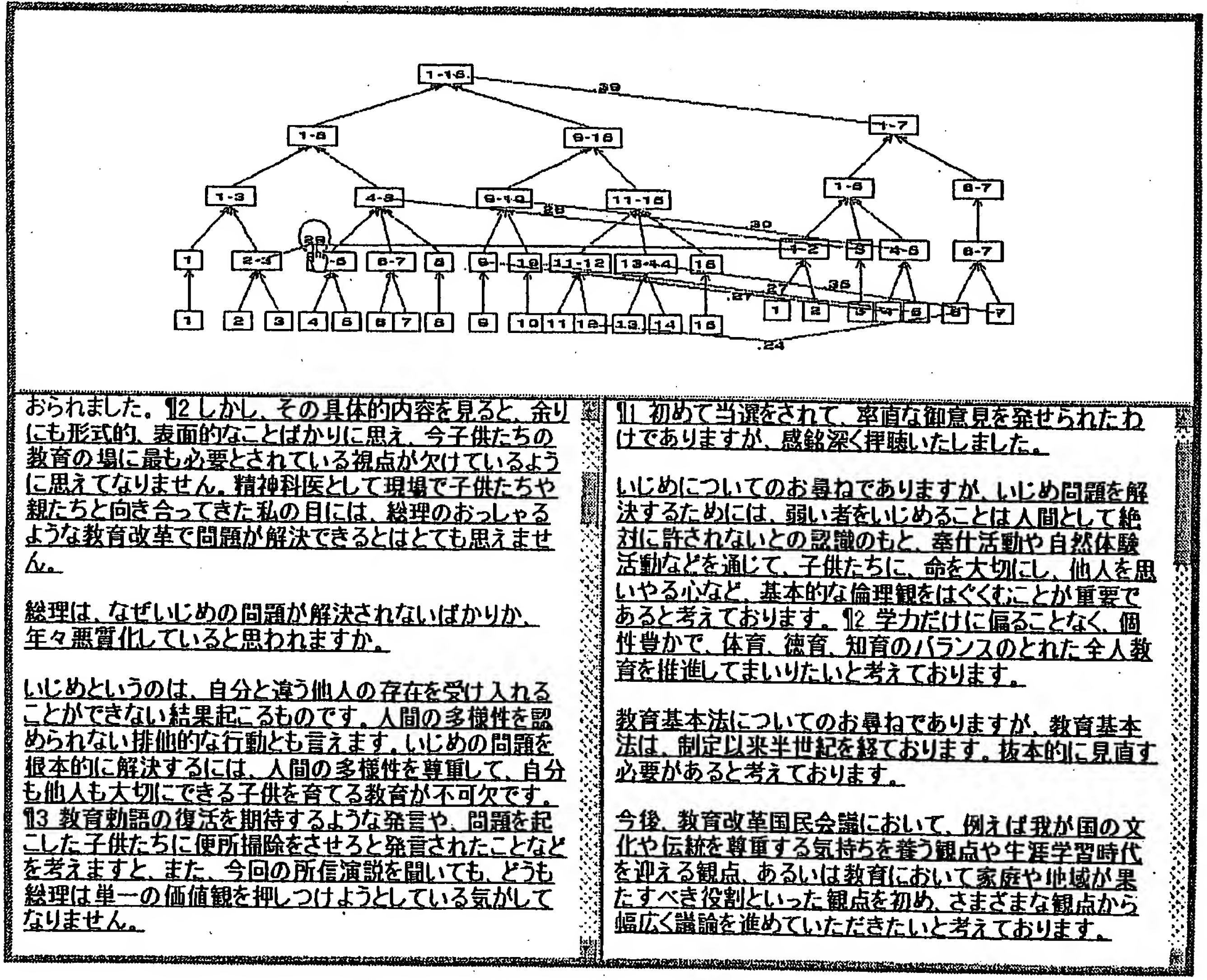
○内閣総理大臣（森喜朗君）

¶1 初めて当選をされて、率直な御意見を発せられたわけでありますが、感銘深く拝聴いたしました。

いじめについてのお尋ねでありますが、いじめ問題を解決するためには、弱い者をいじめることは人間として絶対に許されないとの認識のもと、奉仕活動や自然体験活動などを通じて、子供たちに、命を大切にし、他人を思いやる心など、基本的な倫理観をはぐくむことが重要であると考えております。¶2 学力だけに偏ることなく、個性豊かで、体育、徳育、知育のバランスのとれた全人教育を推進してまいりたいと考えております。

教育基本法についてのお尋ねでありますが、教育基本法は、制定以来半世紀を経ております。抜本的に見直す必要があると考えております。

[FIG. 32] EXAMPLE OF THE PRESENTED RELATED PARTS WITH A GRAPH



おられました。¶2 しかし、その具体的内容を見ると、余りにも形式的、表面的なことばかりに思え、今子供たちの教育の場に最も必要とされている視点が欠けているように思えてなりません。精神科医として現場で子供たちや親たちと向き合ってきた私の目には、総理のおっしゃるような教育改革で問題が解決できるとはとても思えません。

総理は、なぜいじめの問題が解決されないばかりか、年々悪質化していると思われますか。

いじめというのは、自分と違う他人の存在を受け入れることができない結果起こるものです。人間の多様性を認められない排他的な行動とも言えます。いじめの問題を根本的に解決するには、人間の多様性を尊重して、自分も他人も大切にできる子供を育てる教育が不可欠です。¶3 教育勅語の復活を期待するような発言や、問題を起こした子供たちに便所掃除をさせろと発言されたことなどを考えますと、また、今回の所信演説を聞いても、どうも総理は単一の価値観を押しつけようとしている気がしてなりません。

¶1 初めて当選をされて、率直な御意見を発せられたわけでありますが、感銘深く拝聴いたしました。

いじめについてのお尋ねでありますが、いじめ問題を解決するためには、弱い者をいじめることは人間として絶対に許されないとの認識のもと、奉仕活動や自然体験活動などを通じて、子供たちに、命を大切にし、他人を思いやる心など、基本的な倫理観をはぐくむことが重要であると考えております。¶2 学力だけに偏ることなく、個性豊かで、体育、徳育、知育のバランスのとれた全人教育を推進してまいりたいと考えております。

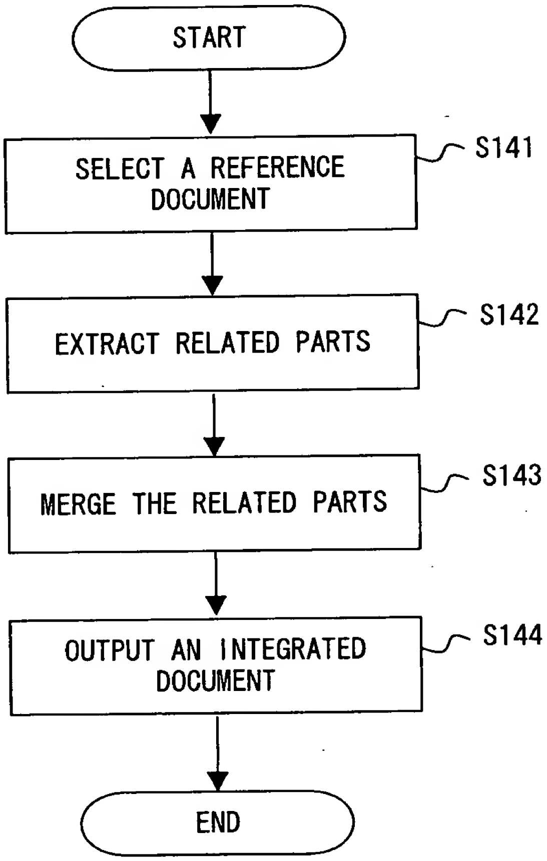
教育基本法についてのお尋ねでありますが、教育基本法は、制定以来半世紀を経ております。抜本的に見直す必要があると考えております。

今後、教育改革国民会議において、例えば我が国の文化や伝統を尊重する気持ちを養う観点や生涯学習時代を迎える観点、あるいは教育において家庭や地域が果たすべき役割といった観点を初め、さまざまな観点から幅広く議論を進めていただきたいと考えております。

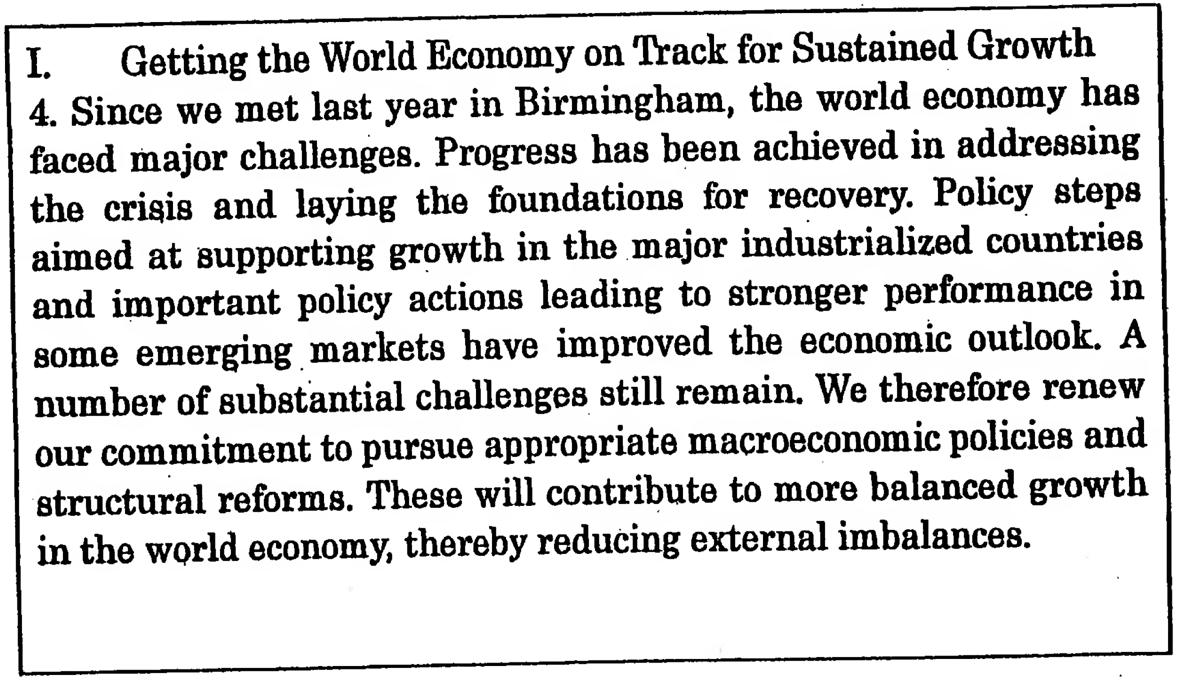
[FIG. 33] EXAMPLE OF THE OUTPUT RELATED PARTS OF THREE DOCUMENTS

○内閣総理大臣（森喜朗君）[22-24]	○水島広子君 [1-3]	○土井たか子君 [17-22]
日本新生プランの第三の柱は、教育の新生、すなわち教育改革であります。悪質な少年犯罪の続発や不登校、学級崩壊などの深刻化は、まことに心痛むものがあります。…命を大切にし、他人を思いやる心、奉仕の精神、日本の文化、伝統を尊重し、国や地域を愛する気持ちをはぐくみ、二十一世紀の日本を支える子供たちが、創造性豊かな立派な人間として成長することこそが、心の豊かな美しい国家の礎と言えるのではないでしょうか。私は、かねてから体育、徳育、知育のバランスのとれた全人教育を充実するとともに、世界に通用する技術、能力を備えた人材を育成するため、世界トップレベルの教育水準の確保が必要であると考えてきました。阪神・淡路大震災やナホトカ号重油流出事故のとき、全国津々浦々から若者たちが集まり、献身的にボランティア活動をしていた姿を見て、さすが日本の若者と感動したことを思い出します。…また、制定して半世紀となる教育基本法についても、抜本的に見直す必要があると考えております。教育改革国民会議においても、九月の中間報告に向けて、我が国の教育各般にわたり議論が行われているところであります。私は、学校の運営体制を整備するとともに、教師が、人間が人間を教えるというとうとい使命感に燃えて教育に携わることが何よりも大切であり、ＩＴ教育や中高一貫の教育の推進、大学九月入学の推進、教員や学校の評価システムの導入、教育委員会のあり方なども重要な課題であると考えております。	…少年犯罪についても、加害者に対する更生システムを専門化し徹底すると同時に、被害者のケアを充実するといった課題に目を向けずに、少年法を改正することで安易に厳罰化を図ろうとするような政治の姿勢には大きな危惧を抱いております。総理は、所信表明演説の中で教育の新生について述べておられました。しかし、その具体的内容を見ると、余りにも形式的、表面的なことばかりに思え、今子供たちの教育の場に最も必要とされている視点が欠けているように思えてなりません。…いじめの問題を根本的に解決するには、人間の多様性を尊重して、自分も他人も大切にできる子供を育てる教育が不可欠です。…教育基本法の見直しにしても、本来、他者との触れ合いを通して自発的に育てるはずの奉仕の精神や道徳心といったものを法改正によって一方的に押しつけようとするのであれば、逆効果となり、取り返しがつかないことになると思います。…	…まさに全国民的な課題と言ってよいでしょう。…むしろ、教育基本法改正や少年法改正の論議に深く踏み込むことによって問題は放置されてしまうと思えてなりません。…子供たちの荒れるのも、クラスの崩壊も、暴力も、子供たちが人間として尊重されていないところから始まっているように私には思えます。自分自身の命と人生を何よりも尊重され、大切に思える子供は、他人の命と人生も尊重し、大切にするものです。…奉仕活動を学校で強制するなど、無意味どころか反発を呼んで逆効果であろうと私は思います。…ボランティアの強制など、言葉の矛盾であるばかりでなく、ようやく日本にも育ちつつある若者たちの本来のボランティア精神をも押し殺してしまうことになるでしょう。…総理、あなたは所信表明で、教育基本法を抜本的に見直す必要を言われました。…教育基本法の改正を提起しようとしている教育改革国民会議についてもお尋ねいたします。…

[FIG. 34] FLOWCHART SHOWING THE SENTENCE INTEGRATION PROCESS



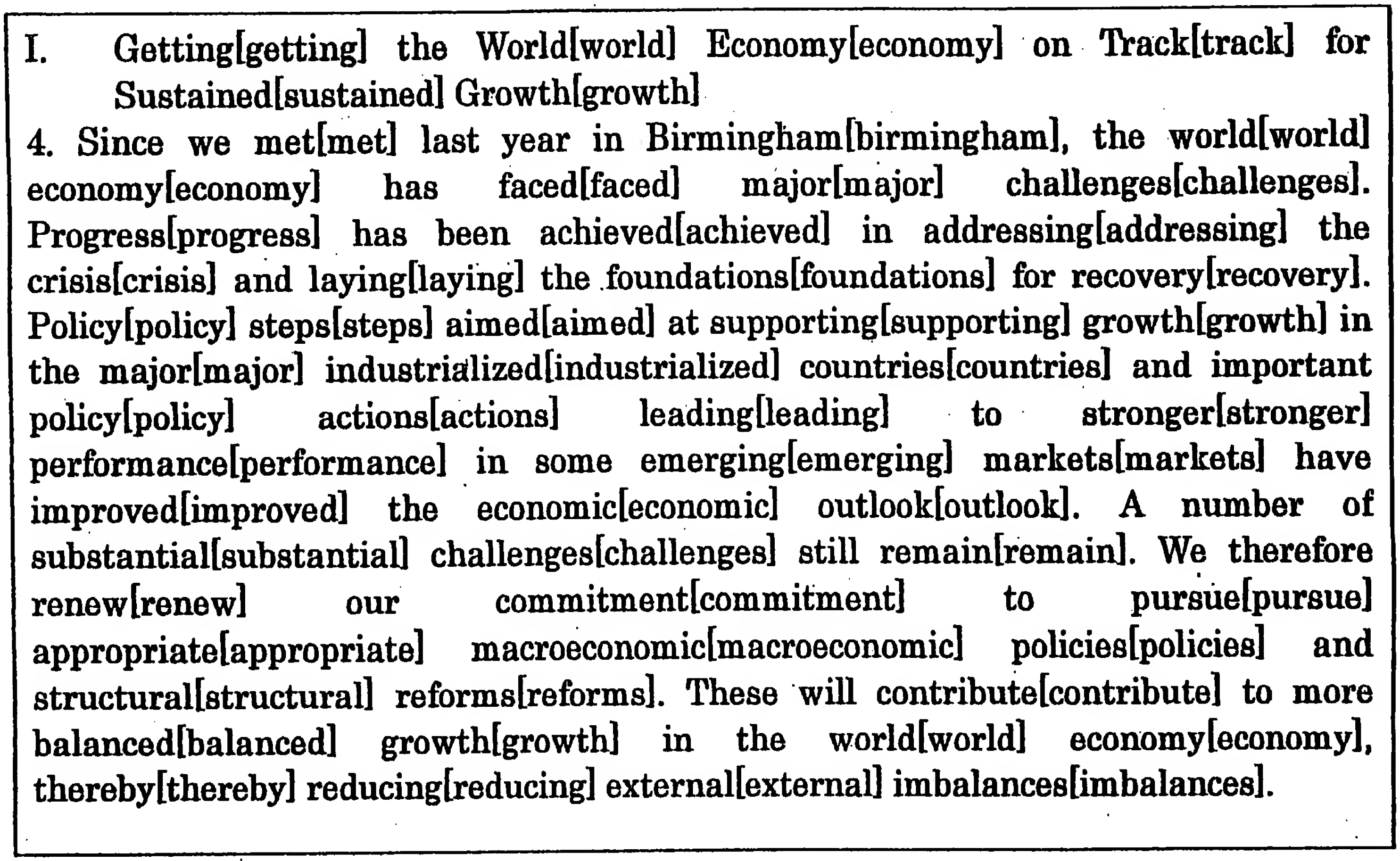
[FIG. 35] EXAMPLE OF THE LEADING PART OF AN ENGLISH DOCUMENT TO BE READ



I. Getting the World Economy on Track for Sustained Growth

4. Since we met last year in Birmingham, the world economy has faced major challenges. Progress has been achieved in addressing the crisis and laying the foundations for recovery. Policy steps aimed at supporting growth in the major industrialized countries and important policy actions leading to stronger performance in some emerging markets have improved the economic outlook. A number of substantial challenges still remain. We therefore renew our commitment to pursue appropriate macroeconomic policies and structural reforms. These will contribute to more balanced growth in the world economy, thereby reducing external imbalances.

[FIG. 36] EXAMPLE OF THE WORD RECOGNITION RESULT OF AN ENGLISH DOCUMENT TO BE READ



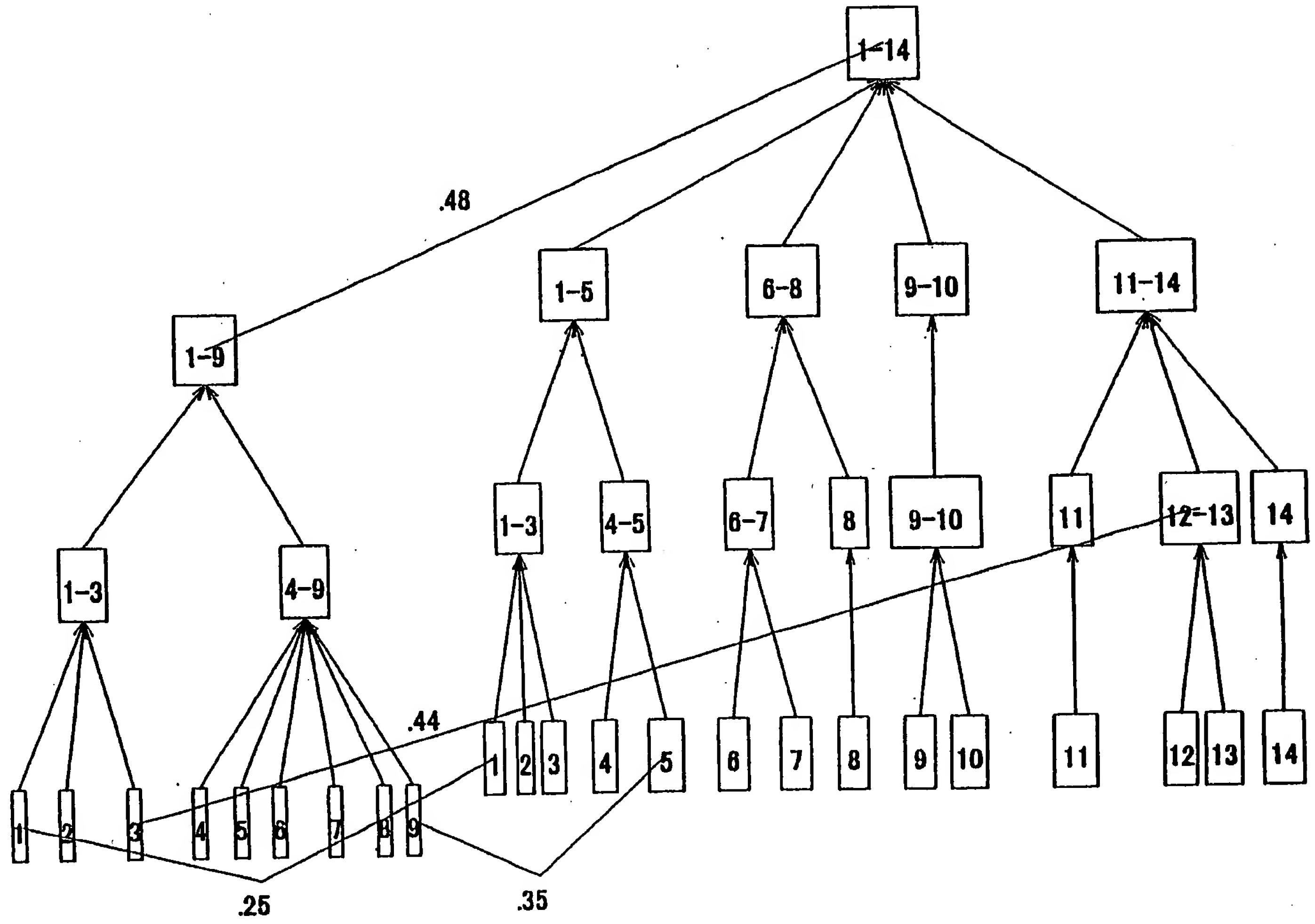
I. Getting[getting] the World[world] Economy[economy] on Track[track] for Sustained[sustained] Growth[growth]

4. Since we met[met] last year in Birmingham[birmingham], the world[world] economy[economy] has faced[faced] major[major] challenges[challenges]. Progress[progress] has been achieved[achieved] in addressing[addressing] the crisis[crisis] and laying[laying] the foundations[foundations] for recovery[recovery]. Policy[policy] steps[steps] aimed[aimed] at supporting[supporting] growth[growth] in the major[major] industrialized[industrialized] countries[countries] and important policy[policy] actions[actions] leading[leading] to stronger[stronger] performance[performance] in some emerging[emerging] markets[markets] have improved[improved] the economic[economic] outlook[outlook]. A number of substantial[substantial] challenges[challenges] still remain[remain]. We therefore renew[renew] our commitment[commitment] to pursue[pursue] appropriate[appropriate] macroeconomic[macroeconomic] policies[policies] and structural[structural] reforms[reforms]. These will contribute[contribute] to more balanced[balanced] growth[growth] in the world[world] economy[economy], thereby[thereby] reducing[reducing] external[external] imbalances[imbalances].

[FIG. 37] EXAMPLE OF A STOP WORD

about, across, again, against, all, along, already, also, among, an, and, another, are, area, areas, around, as, asked, asking, at, back, be, because, been, before, behind, being, best, better, between, both, but, by, can, certain, clearly, come, different, do, during, each, early, end, enough, even, every, everyone, face, fact, few, first, for, from, fully, further, general, gets, give, given, go, good, goods, greater, groups, has, have, having, high, higher, how, if, important, in, interest, interesting, interests, into, is, it, its, keep, know, last, later, least, less, longer, made, make, many, member, members, more, most, much, must, necessary, need, needed, needs, new, next, no, not, now, number, of, older, on, once, one, only, open, opens, order, other, others, our, out, over, part, place, point, points, possible, present, presented, problem, problems, put, right, same, shall, should, shows, since, small, so, some, states, still, such, take, taken, than, that, the, their, them, then, there, therefore, these, they, this, those, three, through, thus, to, together, too, toward, two, under, up, upon, us, use, was, way, ways, we, well, what, when, where, which, while, who, will, with, within, without, work, working, works, would, year, years, yet, young

[FIG. 38] EXTRACTION RESULT OF AN ENGLISH COMMON TOPIC



[FIG. 39] EXAMPLE OF THE SUMMARIZED ENGLISH RELATED PARTS NO. 1)

KÖLN 1999 [1]	**OKINAWA 2000 [1]**
... Progress has been achieved in addressing the crisis and laying the foundations for **recovery**. ... The **world economy** is still feeling the effects of the financial **crises** that started in Asia two years ago. ...	... Yet the financial and economic **crises** of the past few years have presented enormous **challenges for the world economy**. Together with many of our partners around the **world**, we have devoted ourselves to alleviating the adverse effects of the crisis, stimulating economic **recovery**, and identifying ways to help prevent future upheavals, including measures to strengthen the international financial architecture: ...

[FIG. 40] EXAMPLE OF THE SUMMARIZED ENGLISH RELATED PARTS NO. 2)

KÖLN 1999 [3]	**OKINAWA 2000 [12-13]**
II. Building a World Trading System That Works for Everyone The multilateral trading **system** incorporated in the World Trade Organization (**WTO**) has been key to promoting international **trade** and investment and to increasing economic growth, employment and social progress. ... Given the **WTO's** vital **role**, we agree on the importance of improving its **transparency** to make it more responsive to civil society while preserving its government-to-government nature. ... We therefore call on all nations to launch at the **WTO** Ministerial Conference in Seattle in December 1999 a new round of broad-based and ambitious **negotiations** with the aim of achieving substantial and manageable results. ... In this context we reaffirm our commitment made in Birmingham last year to the least developed countries on improved **market** access. We also urge greater cooperation and policy **coherence** among international financial, economic, labor and environmental organizations. ...	... The adoption of the short-term package in Geneva, regarding implementation of Uruguay Round undertakings, increased **market** access for the LDCs, technical assistance for enhanced capacity **building)**as well as improvement in **WTO** transparency, was an important first step in this direction and must be pursued expeditiously. ... We must ensure that the multilateral trading **system** is strengthened and continues to play its vital **role** in the **world** economy. ... We agree that the objective of such **negotiations** should be to enhance **market** access, develop and strengthen **WTO** rules and disciplines, support developing countries in achieving economic growth and integration into the global trading **system**, and ensure that **trade** and social policies, and **trade** and environmental policies are compatible and mutually supportive. ... In this regard, international and domestic policy **coherence** should be enhanced, and co-operation between the international institutions should be improved. ...

FIG. 40

[FIG. 41] EXAMPLE OF THE SUMMARIZED ENGLISH RELATED PARTS NOP. 3)

KÖLN 1999 [9]	**OKINAWA 2000 [5]**
... We also welcome the increasing cooperation between the **ILO** and the **IFIs** in promoting adequate social **protection** and core labor **standards**. ... In addition, we stress the importance of effective cooperation between the WTO and the **ILO** on the social **dimensions** of globalization and **trade** liberalization.	... We also welcome the increasing co-operation between the International Labour Organisation (**ILO**) and the International Financial Institutions (**IFIs**) in promoting adequate social **protection** and **core** labour **standards**. ... In addition, we stress the importance of effective co-operation between the World Trade Organisation (WTO) and the **ILO** on the social **dimensions** of globalisation and **trade** liberalisation. ...

FIG. 41

[Document Name] Abstract

[Abstract]

[Object]

It is an object of the Present invention to extract appropriate related parts for different topics in grading included in a plurality of documents and present them.

[Means for Solving the Problems]

The thematic hierarchy recognition unit 1 recognizes the thematic hierarchy of each of a plurality of documents to be read. The common topic extraction unit 2 extracts common topics that commonly appear in the plurality of documents, based on the recognized thematic hierarchy. The common topic presentation unit 3 extracts description parts corresponding to the extracted common topics from each document in order to support the comparison work of the plurality of documents, and outputs them as related parts,

[Selected Drawing] Fig. 1